DOCUMENT RESUME

ED 238 945 TM 840 051

AUTHOR	Baker, Eva L.; Herman, Joan L.
TITLE	Testing in the Nation's Schools: Collected Papers. Research Into Practice Project.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Nov 83
GRANT	NIE-G-83-0001
NOTE	238p.; Papers presented at the Paths to Excellence: Testing and Technology Conference (Los Angeles, CA. July 14-15, 1983). For related document, see TM 840 043.
PUB TYPE	Collected Works - Conference Proceedings (021) -- Information Analyses (070)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*Achievement Tests; Administrator Attitudes; *Data Analysis; Educational Policy; *Educational Testing; Elementary Secondary Education; Ethnography; National Surveys; *Public Schools; Teacher Attitudes; Test Results; *Test Use
IDENTIFIERS	*Data Interpretation

ABSTRACT

The Center for the Study of Evaluation, of the Graduate School of Education at the University of California at Los Angeles (CSE) hosted a two day conference on "Paths to Excellence: Testing and Technology" on July 14-15, 1983. Attended by over 100 educational researchers, practitioners, and policymakers, the first day of the conference focused on issues in educational testing; day two explored the status and future of technology in schools. This document presents the collected papers from the first day of the conference. Presentations focused on CSE's study of teachers' and principals' use of achievement testing in the nation's schools. The study provided basic data about the nature and frequency of classroom testing, the purposes for which test results are used, principals' and teachers' attitudes toward testing, and local contexts supporting the use of tests (e.g. amount of staff development, testing resources, leadership support). The findings were presented at the conference, and presenters were asked to provide their interpretations of the data and their perspectives on their implications for national, state, and/or local testing policies. One speaker, William Coffman, was asked to provide context for the conference by considering the study in the light of the history of research on educational testing. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED238945

RESEARCH INTO PRACTICE PROJECT

TESTING IN THE NATION'S SCHOOLS:
COLLECTED PAPERS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
● Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Eva L. Baker and Joan L. Herman
Project Directors

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. Gray
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Grant Number:
NIE-G-83-0001

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California

November 1983

TM 840 051

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Contents

Introduction..1

Testing in the Schools: A Historical Perspective
William Coffman...3

Achievement Testing in American Public Schools:
A National Perspective
Donald Dorr-Bremme, Joan Herman, and William Doherty..........28

Testing in the Schools: Implications of a National
Survey of Teachers and Principals
Robert Linn...94

Conceptions of Testing in the Public School
Robert Calfee..115

Testing in the School: An Ethnographic Perspective
Harry Wolcott..134

In Tests We Trust? Remarks on the Pattern of
Test Use in Our Schools
Philip Jackson...146

A Los Angeles Unified School District Perspective
Floraline Stephens...165

Testing in the Schools
Francisco Sanchez..173

Testing in the Schools: A Statewide Assessment
Perspective
Dale Carlson...187

Current Technical Needs of Schools
Carl Sewell..206

The Assessment Needs of Teachers and Administrators
Archie La Pointe...218

Introduction

The UCLA Center for the Study of Evaluation (CSE) hosted a two day conference on "Paths to Excellence: Testing and Technology" on July 14-15, 1983. Attended by over 100 educational researchers, practitioners, and policymakers, the first day of the conference focused on issues in educational testing; day two explored the status and future of technology in schools.

This document presents the collected papers from the first day of the conference. Presentations focused on CSE's NIE-funded study of teachers' and principals' use of achievement testing in the nation's schools. The study provided basic data about the nature and frequency of classroom testing, the purposes for which test results are used, principals' and teachers' attitudes toward testing, and local contexts supporting the use of tests (e.g., amount of staff development, testing resources, leadership support). The findings were presented at the conference, and presenters were asked to respond to them by providing their interpretations of the data and their perspectives on their implications for national, state, and/or local testing policies. Specifically, speakers were asked to do the following:

I. Identify an important question or area of concern in testing and/or education.
II. Discuss the findings in light of the identified questions.
III. Identify next steps for research and/or policy and practice, e.g.,
    What are the implications for teaching practice?
    What are the implications for test development?
    What are the implications for national policy?
    What are the implications for state policy and practice?
    What are the implications for local policy and practice?

Speakers were chosen to represent a balance of national, state, and local policy perspectives as well as a range of disciplinary vantage points. In addition to presenters directly addressing the study findings, one speaker, William Coffman, was asked to provide context for the conference by considering the study in the light of the history of research on educational testing.

Testing In the Schools:
A Historical Perspective

William E. Coffman
E.F. Lindquist Professor Emeritus
College of Education
University of Iowa

> Teachers are important people. They are the people directly responsible for the education of the children and youth of our country. The curriculum of the school is largely what they make it. The professor of education, the school administrator, or the curriculum director may have a large part in determining the content of printed courses of study. They may be responsible for much of the talking and writing in the field of education. But what goes on in the school depends on the teacher in the classroom--on the way he accepts and implements the ideas of the experts or adds his own creative touch based on his unique experience with a particular group of pupils. The teacher, then, is a key person in any program of curriculum development. (Coffman, 1951, p. 305)

I wrote these words a long time ago and in a context different from that of today's conference. But I believe that with a little modification they can be made relevant to the topic of testing in the schools today. Teachers are indeed important people, not only in determining the actual curriculum but also in determining how tests are used in relation to teaching and learning. The legislator, in Washington or the state capitol, may pass laws that mandate specific testing programs; school administrators, in the Department of Education of the nation or state, or of the local school system, may publish edicts or require periodic reports; experts in educational and psychological measurement may argue issues, collect data and publish interpetation, and admonish teachers to do this or that; but, at least in most educational settings, what actually happens is determined by teachers as they interact with pupils in classrooms. One might, therefore, with good reason, ask why it is that so little hard data are available on what actually does happen. And if one wants to make sense of the limited data that are in hand, how must they be organized and interpreted?

I found myself searching my own professional experience for answers to these questions, and then checkng my impressions by referring to more than a half century of published literature. The year I made the decision to enter the field of education, 1931, was the first year of publication of the Review of Educational Research; and two years later the February issue provided the first review on the topic "Educational Tests and Their Uses", a review that cited 467 references (Wood, 1933). The Education Index first appeared in 1929, and the first bound volume in the University of Iowa library (January 1929-June 1932) contains entries under the headings "Examinations" and "Tests and Scales" that reflect interest in and concern with issues still of relevance today: "Examinations as an aid to learning"

(Jersild, 1929), "Examinations seventy-five years ago and today" (Fish, 1930), "Conflicting philosophies concerning educational measurement" (Brown, 1931), "History of the measurement movement" (Malin, 1930), and "Participation in testing programs by the classroom teacher" (Macken, 1929). The heading "Evaluation" first appeared in the next bound volume (July 1932-June 1935), but there was only one entry. Entries increased rapidly during the late 1930's and through the 1940's as concerns broadened to educational outcomes other than recall of information.

The Review of Educational Research carried reviews concerned with testing in the schools at approximately three-year intervals until a more focused and less comprehensive format was adopted during the 1970's. The Education Index marked the growing complexity of the field by expanding the variety of headings, as did the Encyclopedia of Educational Research, beginning with the first edition in 1941. From time to time, the National Society for the Study of Education focused on research and testing in one or another of its yearbooks. And more recently, the annual Review of Research in Education and the ERIC publications have helped us keep on top of a proliferating literature.

The span of my own professional career covers the period since these systematic reviews first appeared in the literature. The first third of the period since then (1931-1949), I was a classroom teacher and administrator in public schools. Since 1949, I have worked as a specialist in measurement and evaluation. The literature, then, serves to confirm, deny, or expand my own recollections.

This is not to say that measurement first became a topic of concern to educators in the 1930's. I note, for example, that the

Twenty-First Annual Conference of Educational Measurement was held at the University of Indiana in 1934, and that Scates was looking back over a period of 50 years as early as 1947 (Scates, 1947). But conferences are often more opportunities for the sharing of impressions than for the reporting of solid evidence, and histories can focus on the highlighting of deficiencies and admonitions for sounder procedures in the future than on the documentation of accomplishments. It was certainly very soon after the accumulated literature began to be systematically reviewed that the scientific movement in education came of age (NSSE, 1935; 1938), and the decade of the 1930's was particularly productive in new insights and challenges. As one of the leaders in the organization of the educational research profession noted at the time,

> Each generation seems to discover for itself teleological and methodological concepts which it brands as new, or progressive, even though these very ideas may have been formulated and voiced centuries or millenniums earlier. It is difficult to know what is new; most ideas are new only to individuals. It appears, however, that there are strong movements in education today which are actually affecting practice in conventional schools in ways which heretofore was only talked about, or practiced in a few private schools. (Scates, 1938, p. 523)

It might be profitable for today's educational researchers, many of whom have brought the conceptual framework and methodological concepts of other academic fields to the study of educational problems, to become acquainted with the educational research literature of the 1930's. The vocabulary may be different, and the total context may be less well-defined than that of today; but the underlying concepts and ideas may often be the same as those that guide today's research.

Themes, Developments, and Cycles

As I have already implied, many of the concepts, issues, and controversies that engage the educational research community today had already been identified early in the 1930's. One can trace these through the literature. In some cases, one finds recurring themes such as a concern with the possibility that standardized tests may have undesirable effects on school curricula. Sometimes there appears to be cyclical movement as a concern shifts from a focus on minimum essentials to a concern with personality development and back again to minimum essentials. In rare instances, one can detect what appears to be real progress, but the progress is more likely to be in a wider dissemination of insights than in the originality of the insight.

For instance, the beginning of concern for efficiency in education through application of principles from business and industry has been attributed to a paper by Franklin Bobbitt in the 12th Yearbook of the National Society for the Study of Education (1913). In that paper he urged careful specification of what pupils were expected to learn in school, and implied that once objectives were specified, teachers might reasonably be held accountable for seeing that they were achieved. One can see the roots of much of today's concern about minimum essentials in the writing of disciples of

Bobbitt over the years. But disciples seldom encompass the full vision of the master, and it is instructive to read what Bobbitt had to say about the importance of considering higher as well as lower level objectives:

> The higher, however, must (also) be scaled. However difficult it may seem to set up quantitative standards in the more intangible field, it must of necessity be done, if once they are introduced into the lower, more objective and more mechanical forms of training. It will work harm to establish definite standards for only a portion of education, leaving the rest to traditional vagueness and uncertainty of aim...But education must take care of all desirable aspects of human personality--training and developing each in due porportion, slighting nothing, neglecting nothing, giving unduly large or unduly small attention to nothing. (p. 26)

Bobbit recognized that it wouldn't be easy to quantify the intangible objectives, and the concern he expressed is still with us today. Much of the controversy over educational measurement in the schools since that time has been concerned with the effect of imbalance in the use of tests, and people are still trying to provide measures of higher level outcomes to redress the balance.

As one prepares to look at testing practices in the schools of the 1980's, it will be profitable to review briefly some of these trends over the years, and to consider their implications for interpeting what we see. Let us begin by considering what we know about teachers' preparation for using tests.

Teacher Education in Testing

At the time that I completed my undergraduate program in secondary education, my home state of West Virginia required that all applicants for certification as a teacher in the secondary schools had completed a course in tests and measurement. I was enrolled in a college in Ohio, and since Ohio did not have such a requirement, I completed the requirement through individual study. At the time, the fact that such a requirement was not widespread was of little significance to me; but what about now? Apparently, the passing years have not seen much change in the situation. At mid-century, Betts (1950) was taking a dim view of the ability of teachers to interpret standardized test results:

> Such norms (GE) are highly satisfactory to teachers because pupils in general make greater progress during the course of the year than is shown in cross-sectional norms. When standardized testing is done at the beginning of the school year, teachers using the test find a majority of their pupils above the norm at the end of the school year and glow with success. They are unaware that the test they are using probably measures intelligence, not school taught learnings, and that what appears to be greater than normal progress, is a mere

statistical artifact. (p. 218)

In 1959, Mayo reported a study by Noll indicating that 83% of 80 colleges he had surveyed offered a course in measurement, but that only 14% of them required one of all teacher education students. Furthermore, only 10% of the states required a course for certification. Ten years later Stinnet (1969) made no mention of any requirement in educational measurement in his encyclopedia article on teacher certification, nor did Burdin (1982) thirteen years later. It seems obvious that only a minority of teachers have had any intensive training in educational measurement. Is it possible that those who have may exhibit quite different practices from those who have not? Certainly, information regarding the background in educational measurement of respondents would appear to be critical in the interpretation of survey reponses.

To those of us in the measurement profession, the lack of course work in the field in programs of teacher education appears to be a serious omission. The fact that it apparently does not seem so to other educators suggests a need to look more closely. What does such a look reveal?

Teachers and Researchers

One thread running through the measurement and evaluation literature is a concern, on the part of measurement specialists, that teachers seem not to be taking seriously the admonitions of researchers and measurement specialists regarding ways of using tests in classroom settings. The concern seems seldom to have led to the collection of hard data. One explanation for this phenomenon may be found in an analysis of the problem by Scates (1943). Scates pointed out that the scientist is interested in truth leading to broad generalizations, while the teacher seeks information of direct practical value; the scientist is interested in elements, whereas the teacher is interested in functioning organisms; the measurement specialist cannot measure continuously, but the teacher needs to and must measure continuously; the scientist measures traits uniform throughout their range, but the teacher measures growth in stages; and the measurement specialist generally measures formal abilities by cross-sectional power tests, but the teacher must be concerned with behavioral dynamics in life situations.

To the extent that Scates's analysis is sound, it is not suprising that there is little systematic study of teachers' testing practices reported in the literature written primarily by researchers and test specialists. They had their own interests, which were different from those of teachers, and they probably weren't even aware that the difference existed.

It is true that over the years the interests of researchers have turned more from concern with simple elements to concern for the dynamics of learning. Still, recent articles tend to confirm the conclusions of Scates:

> Teacher preference, in effect, is for continuous movies, in color with sound, while a test score, or even a profile of scores, is more akin to a black-and-white photograph. (Salmon-Cox, 1981)

There is even a tendency to focus on uses of tests in research and guidance rather than as tools in the instructional setting. For example,

> Two functions of tests that deserve particular emphasis at this time are: first, the uses of educational tests in the construction and evaluation of educational theories, especially theories that give particular attention to processes or strategies of problem-solving rather than outcomes alone; and second, the uses of tests in the service of individual students through systems of guidance that employ measurement as a means of fostering self-discovery and as a means for encouraging students to develop wisdom in decision-making. (Manning, 1970, pp. 20-21)

To some extent, recent interest in qualitative methods have brought the data collection procedures of the researcher closer to the interests of the teacher (Hamilton et al., 1977). But it is unlikely that teachers generally will seek greater expertise in anthropological methods than they have in psychometric methods. It is more likely that if they wish to increase the use of tests in instructional settings, researchers will need to be asking themselves: what is it in our materials and methods that is likely to be useful to teachers whose basic guides to decisions are the moment-by-moment observations so clearly described by Jackson (1968) in Life In Classrooms. And the researcher interested in how teachers use tests will want to collect enough information about the total mix of data, observation as well as formal and informal, testing to understand the place of testing in the mix.

Incidentally, it appears that often the teacher's orientation is different, not only from that of the researcher and test specialist, but also from that of the school administrator and school board member. This idea is well expressed by Gorton (1982, p. 1906):

> Teachers tend to emphasize such aspects as humanistic orientation to instruction and positive relations between teachers and students; administrators, on the other hand, stressed such factors as student achievement on standardized tests and administrative evaluation.

Given that such differences do exist (the research tends to be based on small and often non-representative samples), recent trends toward differentiation of testing in relation to function would probably be welcomed by teachers. Lefever (1950) expressed the possibilities quite clearly almost 25 years ago. He argues (but with no supporting data) that teacher-made tests should be considered essential tools for checking pupil achievement, particularly at the secondary school level; that teachers grow in professional competence as they participate in test construction; that specialists in measurement should be active in in-service education to facilitate sound teacher activity; that general survey testing to evaluate educational programs should never be broken down to the individual class level and might well be conducted using matrix sampling; and that it is essential for teachers to be actively involved in planning the system testing program. To the extent that separation of function of this sort is operating, responses of teachers to survey questions may be expected to differ from those under different circumstances.

Different Philosophical Positions

Another issue that has complicated the picture of testing in the schools involves much more than differences between teachers and test specialists, or between teachers and administrators. In fact, there is almost never a simple contrast, for within each of these groups there are likely to be differences about the purposes of education, the nature of human learning, and the nature of evidence, that is, differences in basic philosophy (Coffman, no date; Hughes, 1934; Thelen, 1969; Weiss, 1981). While the proportions of each group holding a particular position may vary, all positions are likely to be found within each group. Furthermore, the philosophical domain is not a simple one that can be represented by a single dimension, for example, conservative-liberal. In most cases, one needs to look for various dimensions.

There is, for example, the issue of whether the school should be concerned primarily with the transmission of the culture to each new generation or primarily with the development of skills needed for adjusting to a constantly changing culture. There seems little doubt that Bobbitt (1913) was concerned primarily with the former, although his view of the culture to be transmitted was broader than that of many of his followers. Findley and Smith (1950, p. 63) called attention to a contrasting position argued by Brownell (1948). They wrote:

> Brownell offered a criticism of learning implicit in most educational measurement. He insisted that we raise our sights from measures

> of rate and accuracy of performance to measures of level of process used, from evidence of immediate gains to that of more permanent gains, and from ability to use learning in closely similar situations to transferability to essentially new situations, especially after a significant lapse of time.

More than a decade earlier, Brownell (1937, p. 492) had posed a challenge to test developers that is still challenging them today:

> To meet the proposed criteria, a test must (1) elicit from pupils the desired types of mental process, (2) enable the teacher to observe and analyze the thought processes which lie back of the pupils' answers, (3) encourage the development of desired study habits, (4) lead to improved instructional practice, and (5) foster wholesome relationships between teacher and pupils.

Snow, writing in 1980, sounds the same note, but perhaps the tools for tackling the problem are more appropriate than they were in 1937.

> If one looks only at immediate achievement, ignoring aptitude, and most instructional research still does both of these things, then elaboration of instruction appears beneficial. If one adds general ability to the picture, it turns out that elaboration helps the less able

> learners but may not be optimal for the most able learners. If one must further choose a particular form of elaboration to give less able students, it appear best to match the form to the learner's relative strengths. However, when retention is considered, all this changes. Unelaborated instruction is best for almost everybody, and particularly for students high in verbal-crystallized ability. And if one had to choose a form of elaboration, it would seem best to mismatch the form with a student's ability profile.(p. 56)

Other researchers and test specialists are also showing an interest in the development of tests that can provide data directly applicable to issues in testing and learning (Anderson, 1972; Calfee, 1981; Messick, 1983). In each case, however, the concern is with education designed to develop intellectual skills rather than to transmit information. To teachers who accept the skills objectives, the message in the literature is likely to be significant. To those whose orientation is toward content as the focus of education, the message may have little impact. And what about those holding other positions: that the purpose of education is the cultivation of well-adjusted, happy individuals, or the building of a new social order?

The concern with personality development that characterized the progressive education movement in the 1930's does not seem to be of much concern to researchers and testers today, but there are undoubtedly many with roots in this position who occupy teaching positions today and whose philosophical orientation leads them to the view that tests that focus only on either information or intellectual skills are restrictive. To them, the methods of the clinician are preferable to those of the psychometrician, and their responses to questions about testing and evaluation will make sense only when the philosophical context is made explicit. They might, however, be surprised to read this quotation from Wood's article in the Review of Educational Research in 1933:

> ...the highest purpose and ultimate aim of the objective testing movement is not to make better college entrance or course-credit examinations, but to help inaugurate a continuous study of individuals throughout the whole educational ladder by means of systematically recorded comparable measures and observations which will make such spasmodic examinations largely unnecessary...The first question that the school should ask and answer at least provisionally several times a year is, "What can Johnny learn, and which of the things he can learn should the school, in the light of all the facts, try to help him learn?" Tests should first of all tell what a pupil should try to learn--not how he may be cajoled, persuaded, or insidiously coerced into the learning item *x* in the "standard" curriculum for grade *n*. (pp. 7-9)

Testing and Public Policy

One factor that may well influence the reactions of teachers to test and evaluation practices, and so be critical to the interpretation of research concerned with the use of tests, is the extent to which policy decisions by public agencies depend on test results. Traditionally, in the United States, policy decisions regarding schooling have rested in the hands of local agencies, and for such decisions, little use has been made of formal testing. In the continuing discussion of ways in which tests might influence teaching practices, there has been recognition of the need to guard against giving too much weight to test results. In fact, as early as the mid-1930's, when Lindquist was establishing the Basic Skills Testing Program in Iowa, he cautioned that test results, if they were to be useful in guiding teaching and learning, should not be used for the purpose of evaluating teachers or for rating schools (Peterson, 1983). Early studies of teacher practices and attitudes were carried out in this context, and interpetations of results even as late as 1981 may be reflecting to a certain extent the tradition of local control and autonomy. Miller (1963) indicated that in spite of claims to the contrary, there was little likelihood that state or national testing programs would influence very much the practices of good teachers in the secondary schools. Goslin (1967) reported that many teachers look on tests as of peripheral importance. Salmon-Cox (1981) reported that teachers prefer to depend on their own judgment rather than on test results. However, these studies represent another time--or were based on highly specialized samples. The possible effects of recent trends was clearly recognized by Madaus (1981), who wrote:

> U.S. education is now adopting a new relationship between testing and policy, and hence between test results and their use. Testing is now being asked to assume a new role, one in which a test mandated by a policy board (often external to the local school district) becomes the administrative device through which a particular educational policy is implemented. The effects of such testing programs on the balance of power between local districts and the agency mandating the test are a direct function of the rewards or sanctions associated with test use. Both history and the contemporary experience of western European countries reveal that, whenever test results become a key element in important decisions that affect individual life chances (e.g., graduation from high school or grade-to-grade promotion, teacher salary or tenure decisions, school certification, or the allocation of funds), the agency that administers the test assumes a great deal of power over the schooling process. When external tests are used in these ways, administrators, teachers, and pupils take the results seriously and modify their behavior and attitudes accordingly. (1981, p. 635)

It would appear, then, that for any clear interpretation of data based on surveys of teacher attitudes and practices with respect to tests and testing, it would be important to assess the extent to which respondents were feeling the effects of the use of tests for implementing policy.

Conclusions

What, then, does a survey of the literature related to testing in education (when filtered through the collected observations of one person over 50 years) suggest to researchers today seeking insights into how teachers collect and interpret data about pupil achievement? Perhaps the most important conclusion is that one can't make much sense out of responses to questions unless they are placed in an appropriate context. Answers to questions will vary, and the meaning of those answers will depend on a variety of factors affecting the respondent. The interesting findings will be the interactions between questions and these factors, not the first order responses. More specifically, this review suggests that the researcher of the 1980's should consider these things:

1. Studies in the past of teachers' use of tests have been of two kinds. There have been intensive studies of small and non-representative samples that provide a rich framework for interpretation but leave the reader with the feeling that what the researcher found may be true of these teachers in these settings, but not necessarily of other teachers in other settings. There have also been large-scale surveys that break down responses along easily identified but not necessarily significant categories such as sex, geographical

region, level of education, or size of school or community. What is needed is information based on a comprehensive and representative sample that can be broken down along meaningful dimensions.

2. One factor that may well moderate teacher attitudes and practices may be the extent of training in principles of measurement and evaluation. The evidence is that teachers with formal course work in measurement and evaluation at the preservice level are a minority, and that inservice programs vary all the way from extensive and profound to superficial or nonexistent. It will certainly be helpful in making sense of responses to have information about the respondents' background in testing.
3. The literature documents the rather dramatic difference in the views of teachers and researchers regarding what tests should provide in the way of information. Thus, researchers should be on guard against framing survey questions that may be significant to them but not necessarily to teachers--or against framing questions that may be perceived differently by teachers than intended by the researcher. Researchers might even consider researching the question of whether or not the continuous observation described by such researchers as Jackson or Salmon-Cox may be providing teachers with more valid data than that provided by any single test, however comprehensive.
4. Even though teachers and researchers, or teachers and administrators, or teachers and laymen, may differ in

general in their attitudes toward testing, there will be, in each situation, philosophical viewpoints that are influencing attitudes and values--and practice. Responses may be different, depending on the philosophy of education of the respondent; and for teachers with the same philosophy of education, responses may differ depending on whether or not that philosophical position is held also by administrators in the system or by officials outside the system who are perceived as holding power over the system. The phenomenal field of the respondent needs to be assessed if responses are to be properly interpreted.

5. Finally, the researcher will need to assess carefully the extent to which the use of tests in the implementation of public policy is having an impact on testing in the schools from which respondents are coming. It is not yet clear whether the increased use of tests for such purposes is a trend that will continue, or whether we are near the peak of a fluctuating cycle. In any case, how the teacher or administrator views the distribution of power may well influence the responses collected by the researcher.

References

Anderson, R.C. (1972). How to construct achievement tests to assess comprehension. _Review of Educational Research_, _42_(2), 145-70.

Betts, G.L. (1950). Better interpretation and use of standardized achievement tests. _Education_, _71_(4), 217-221.

Bobbitt, F. (1913). Some general principles of management applied to the problems of city-school systems. In S.L. Parker (Ed.), _The Twelfth Yearbook of the National Society for the Study of Education_ (Part I). Chicago: University of Chicago Press.

Brown, M.E. (1931). Conflicting philosophies concerning educational measurement. _Kadelpian Review_, _10_, 175-179.

Brownell, W.A. (1937). Some neglected criteria for evaluating classroom tests. _National Elementary Principal_, _16_, 485-492.

Brownell, W.A. (1948). Criteria of learning in educational research. _Journal of Educational Psychology_, _39_, 170-182.

Burdin, J.L. (1982). Teacher certification. In H.E. Mitzel (Ed.), _Encyclopedia of educational research_ (5th ed.). New York: The Free Press.

Calfee, R. (1981). Cognitive psychology and educational practice. In D.E. Berliner (Ed.), _Review of research in education_ (No. 9). Washington DC: American Educational Research Association.

Coffman, W.E. (1951). Teacher morale and curriculum development: a statistical analysis of responses to a reaction inventory. Journal of Experimental Education, 19, 305-332.

Coffman, W.E. Concepts of achievement and proficiency. In P.H. Dubois (Chmn.), Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, no date.

Findley, W.G. & Smith, A.B. Measurement of educational achievement in the schools. Review of Educational Research, 1950, 20(1), 63-75.

Fish, L.J. Examinations seventy-five years ago and today: comparison and results. Excerpts. Elementary School Journal, 1931, 31, 330-332.

Gorton, R.A. Teacher job satisfaction. In H.E. Mitzel (Ed.), Encyclopedia of educational research (Fifth edition). New York: The Free Press, 1982.

Goslin, D.A. Teachers and testing. New York: Russell Sage Foundation, 1967.

Hamilton, D. et al. (Eds.) Beyond the numbers game. Berkeley, California: McCutchan, 1977.

Hughes, W.H. Two educational philosophies--Are they compatible? Nation's Schools, 1934, 14(6), 25-8.

Jackson, P.W. Life in classrooms. London: Holt, Rinehart and Winston, 1968.

Jersild, A.T. Examinations as an aid to learning. Journal of Educational Psychology, 1929, 20, 602-9.

Lefever, D.W. The teacher's role in evaluating pupil achievement. Education, 1950, 71(4), 203-9.

Macken, I.N. Participation in testing programs by the classroom teacher. Educational Administration and Supervision, 1929, 15, 117-126.

Madaus, G.F. (1981). Reactions to the 'Pittsburgh Papers'. Phi Delta Kappan, 62(9), 634-6.

Malin, J.E. (1930). History of the measurement movement. Educational Outlook 4, 72-80.

Manning, W.H. (no date). The functions and uses of educational measurement. In P.H. Dubois (Chmn.), Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service.

Mayo, S.T. (1959). Testing and the use of test results. Review of Educational Research, 29(1), 5-14.

Messick, S. (1983). Abilities and knowledge in educational achievement testing: the assessment of dynamic cognitive structures. In S. Elliott and J. Mitchell, Jr. (Eds.), Buros-Nebraska series on measurement and testing, (Vol. 1), B. Plake (Ed.), Social and technical issues in measurement and testing: Implications for test construction and usage. Hillsdale, New Jersey: Erlbaum.

Miller, D.F. (1963). State and national curriculums and testing programs: Friend or foe of expertness in the classroom? Journal of Secondary Education, 38, 41-47.

National Society for the Study of Education. (1935). Thirty-fourth yearbook: Educational diagnosis. Bloomington, Indiana: Public School Publishing Company.

Peterson, J.J. (1983). Iowa Testing Program: The first half-century. Iowa City, Iowa: University of Iowa Press.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 62(9), 631-4.

Scates, D.E. (1938). The improvement of classroom testing. Review of Educational Research, 8(5), 523-36.

Scates, D.E. (1947). Fifty years of objective measurement and research in education. Journal of Educational Research, 41, 241-64.

Snow, R.E. (1980). Aptitude and achievement. In W.B. Schrader (Ed.), Measuring achievement: Progress over a decade. (Proceedings of the 1969 ETS Invitational Conference) San Francisco: Jossey-Bass.

Stinnett, T.M. (1969). Teacher certification. In R.L. Ebel (Ed.), Encyclopedia of educational research (4th ed.). New York: Macmillan.

Thelen, H.A. (1969). The evaluation of group instruction. In The Sixty-eighth Yearbook of the National Society for the Study of Education, Part II: The Scientific Movement in Education. Chicago: University of Chicago Press.

Twenty-first Annual Conference On Educational Measurements. (1934). Indiana University School of Education Bulletin, 11, 1-93.

Wood, B.D. (Ed.). (1933). Educational tests and their uses. Review of Educational Research, 3(1).

Weiss, J. (1980). Assessing nonconventional outcomes of schooling. In D.C. Berliner (Ed.), Review of research in education (No. 8). Washington DC: American Educational Research Association.

Achievement Testing in American Public Schools
A National Perspective

Donald W. Dorr-Bremme, Joan L. Herman, and William Doherty
Center for the Study of Evaluation
University of California, Los Angeles

The UCLA Center for the Study of Evaluation (CSE) began its Test Use in Schools study just as achievement testing in American schools was becoming the subject of increasing public discussion and debate. Critics had begun to decry the arbitrariness of current testing practices (Baker, 1978). They had indicted tests' validity and attacked them as biased (Perrone, 1978), accused testing of narrowing the curriculum, and questioned the value of testing amidst the changing functions of American education (Tyler, 1978). The quality of available tests had become a matter of controversy (CSE, 1979; The Huron Institute, 1978), and at least one major teachers' organization had called for a moratorium on the use of standardized tests. In response to the critics' challenges, advocates of testing had begun to reassert that current tests can and do serve a variety of important purposes. These proponents maintained, for example, that testing promotes accountability, facilitates more accurate placement and selection decisions, and yields information useful for curricular and instructional improvement.

The testing controversy has continued and the stakes in the debate are high. The nation's investment in school achievement testing is enormous, and the amount and variety of testing continue to grow. Simultaneously, school-board accountability demands, mandates for minimum competency (or proficiency) testing, evaluation requirements for federal, state and local education programs, and a variety of judicial decisions on the responsibilities of public schools have combined to make the quality of testing and test use urgent concerns. These and other factors have fueled the testing controversy.

Yet despite this controversy and the importance of the issues it entails, there has been little information forthcoming on the nature of testing as it is actually conducted and used in the schools. How much testing really goes on? What functions do tests serve in the classroom? How are test results used by teachers and principals? What kinds of tests do principals and teachers trust and rely upon most? These and similar questions have gone largely unaddressed. A few studies have indicated teachers' circumspect attitudes toward and limited use of one type of achievement measure -- the norm-referenced, standardized test (e.g., Airasian, 1979; Boyd, et al., 1975; Goslin, 1965; Epstein and Hilloch, 1965; Resnick, 1981; Salmon-Cox, 1981; Stetz and Beck, 1979). Beyond this, however, the landscape of testing practices and test uses in American schools has remained unexplored.

In this context, CSE's three-year study provides basic, new information on classroom achievement testing across the United States. Conducted from 1979 through 1982 (with some data analyses still underway), CSE's research proceeded from broad definitions of test and testing. It encompassed a wide range of types of formal assessment measures (e.g., commercially produced norm- and criterion-referenced tests and curriculum-embedded measures; tests of minimum competency or functional literacy; district-, school-, and teacher-developed tests); as well as some less formal means for gauging student achievement (i.e., teachers' observations of and interactions with learners). Within this broad field, inquiry focused on achievement assessment practices and uses in reading/English and in mathematics as carried out in public schools at the upper-elementary and high-school levels, i.e., in grades 4-6 and 10-12. A nation-wide survey of teachers and principals was central to the study, and results of this survey form the basis of the report that follows. The research also included exploratory fieldwork in preparation for the survey and, following the survey, case study inquiry on testing costs. During these phases of the project, interviews were conducted with approximately 100 school-level educators (including 12 principals and 69 teachers) in five school districts across the country. Interview results were completely consonant with survey findings and yielded a deeper understanding of them. While these interview findings are not presented in detail here,they have influenced the interpretation and discussion of the survey results.

Below, we first provide a brief description of the survey sample, then continue with survey findings on five major questions:

1. How much and what kinds of achievement testing take place in the nation's schools?
2. How important are the results of different types of assessment in teachers' and principals' routine tasks?
3. What are schools' and districts' administrative practices with regard to testing and test use?
4. What are teachers' and principals' perceptions of testing and test use?
5. What factors seem to influence testing practices?

The Survey Sample*

The survey addressed a nation-wide sample of principals and teachers drawn through a successive, random-selection procedure. First, a nationally representative probability sample of 114 school districts was drawn, stratified on the basis of district size, minimum competency testing policy, socioeconomic status, urban-suburban-rural locale, and geographic region of the country. (A lattice sampling technique was used to select cells from the matrix defined by these five stratifying variables, and then random sampling to select districts within a cell.) Next, from within these districts, size permitting, two elementary schools and two high schools were randomly selected using a procedure that facilitated (where possible) inclusion of schools at levels serving both higher- and lower-income populations. Finally, in each of these schools, principals received directions for randomly drawing four teachers for inclusion in the study. (The directions for elementary principals guided the random selection of two fourth-grade and two sixth-grade teachers; those for high school principals, the random selection of two teachers of tenth-grade English and two of tenth-grade mathematics.) The principal and each of the four participating teachers received questionnaires that elicited detailed information on their individual and school testing practices, as well as related contextual and attitudinal data.

---

*A detailed description of the sampling procedure and results is contained in a separate report (Choppin, et. al, 1981). This information has not been reproduced here in order to avoid redundancy. Readers interested in more information regarding the sample and procedure used to draw it are referred to that earlier work.

Returns were obtained from 220 principals, 475 elementary school teachers, and 363 high school teachers in 91 of the 114 districts sampled. Return rates from all principals and from teachers at the elementary level were approximately 60%. About 50% of the high school teachers in the sample responded. To correct for differential return rates by sampling cell and to approximate a nationally representative distribution of respondents, weightings were applied in all descriptive analyses. The results reported below, therefore, represent weighted estimates of national testing practices, test use patterns, and principal and teacher perceptions on testing-related issues.

Before presenting the results derived from the sample described above it would be beneficial to provide some detail about the respondents and their environment. The remainder of this section describes the characteristics of the sample respondents and their schools. Specifically, we will focus on the characteristics of the school context in which the respondents operate and then on the teachers themselves. It is anticipated that this information will help provide a better understanding of the results to be discussed in the later sections.

The typical elementary school in the sample serves a total enrollment of 528, comprised of a majority Caucasian but ethnically mixed student population. While the typical school community is economically heterogeneous, a significant minority of students receive federal aid and/or qualify for free school lunch benefits. Transiency and absence rates are relatively modest, 16 and 6 percent respectively. A majority of the schools (60%) operate a school improvement program, and student achievement testing is typically included and required in such programs. Over one half of the schools operate under minimum competency testing requirements; while within these schools most students pass such required tests on the first try, a sizeable number of students (20%) typically experience failure. (See Table A-1, Appendix).

Secondary school enrollments, as would be expected, are substantially higher, with a mean of 1439. While other characteristics were quite similar to those at elementary school level, students in the average high school in the sample appeared slightly more economcially advantaged, and less transient.

The typical teacher within the schools described above had approximately twelve years of teaching experience, almost ten of which was in their current district. (The results are presented in Table A-2.) In terms of their education the respondents were almost evenly split between a Bachelors and a Masters degree, with less than 1% holding a doctorate. Further, they tended to average some 24 to 25 college units beyond their highest degree. The picture one has, then, of the teachers in the sample is one of an experienced, well qualified professional who has continued to receive education. It is interesting to note how similar the characteristics were across the elementary and secondary levels.

The classroom these teachers tended to operate in is described also in the results found in table A-2 in the Appendix. The results indicate that the teachers had approximately 27 students at the elementary level and 26 at the secondary level. At the elementary level, they provided over 6.5 hours of reading instruction per week and about 5 hours of mathematics instruction. The results at the secondary level were similar for mathematics, i.e., about 5.5 hours of instruction per week. However, fewer hours of English instruction occured at the secondary level than reading instruction at the elementary level. This reflects both the greater emphasis on reading earlier in a student's career as well as the broadening of the curriculum as a student progresses through higher grade levels. It should be instructional to compare these average amounts of weekly instruction with the amount of time devoted to testing, which is described in the following sections.

How Much Testing Goes on in Schools?

Survey results show that the typical student in the upper elementary grades spends, on the average, about 10 hours a year taking reading tests and somewhat more than 12 hours a year taking mathematics tests. (See Table 1.) Test-taking time, then, seems to comprise a little over five percent of the time often allocated annually to formal instruction in each of these subjects. (This figure assumes one hour of daily instruction in each subject for 177 school days per year.)

The typical tenth-grade student enrolled in English, survey results indicate, spends about 26 hours a year completing English tests. This constitutes in the neighborhood of twenty percent of his or her annual time in English class. For the typical tenth grader enrolled in mathematics, taking math tests consumes a little over 24 hours each year -- roughly eighteen percent of the time spent annually in mathematics class. (Here, the percentages given assume daily classes of 45 minutes in each subject, over 177 days per school year.) Clearly, on the average nationally, the frequency and duration of testing in the high school subjects exceed those in the equivalent upper-elementary-school subjects. (Refer again to Table 1.)

It bears reiterating that the annual times on testing reported here are estimates of students' test-taking times. As such, they can probably only serve as rough indicators of the times that the teachers in question spend giving tests in the classroom. On-site interviews (Dorr-Bremme, 1982) suggest that elementary teachers spend only about a quarter to a third of their total time on testing actually giving tests in the class-room. That is, for each hour they devote to giving a reading or math test, they typically spend another two or three hours in such activities as preparing for testing (e.g., constructing and dittoing the test, reviewing directions for standardized testing), correcting and grading tests (or checking over students' standardized-test answer sheets), recording scores, etc. (Time spent consulting test results and otherwise "using" them is not included here.) Thus, elementary-school teachers' annual time on testing far exceeds the typical student's. (Case studies in two elementary schools found that teachers spent on the average of 200 to 250 hours per year, in and out of class, in achievement testing in all subject areas--or roughly 12 to 15 percent of their reported annual work time.) Resources were not

Table 1

Time Devoted to Testing in Typical Classes

	Total Amount of Class Time Spent on Testing per Annum	No. of Test Sessions for Typical Student	Average Length of Session
Elementary School (Grades 4-6)			
--Reading Tests	9 hrs. 56 min.	22	27 min.
--Mathematics Tests	12 hrs. 28 min.	23	32 min.
10th Grade English Class	26 hrs. 34 min.	49	32 min.
10th Grade Mathematics Class	24 hrs. 18 min.	45	33 min.

Table 2

Time Devoted to Required Testing, As a Percentage of Total Testing Time For Typical Classes

	Percentage Time on Testing Required by State	Percentage Time on Testing Required by Local School District	Percentage Testing Time Devoted to Non-Required Tests
Elementary School (Grades 4-6)			
--Reading	30	29	41
--Mathematics	21	25	54
10th Grade English Class	12	13	74
10th Grade Mathematics Class	9	14	77

available for detailed case studies in high schools, but pre-survey interview data indicate that the average testing time per year of high-school teachers is also much greater than their students'.

How much of the testing just described is required by the educational hierarchy beyond the school? How much is undertaken at the discretion of teachers? Table 2 provides data to answer these questions. Elementary teachers in the sample report that about half the testing they conduct both in reading and in math is required by their state or school district. At the high school level, about one quarter of the classroom assessment in both English and mathematics results from state or school-district mandates. Notice, then, that since high school students on the average spend twice as much time annually being tested as elementary students do, these percentages suggest that the actual number of hours spent in required testing is quite similar at both levels of schooling. Notice, too, that a greater proportion of assessment in the high school subjects is voluntary: conducted at the discretion of the individual teacher.

What types of tests are used most heavily? Which types consume larger proportions of classroom testing time? As Table 3 shows, tests developed by individual teachers and schools and, at the elementary level, those which accompany curriculum materials, occupy the great majority of classroom testing time. Of all the test types listed, these are the types over which teachers have most control. They can administer them when they deem appropriate; they can design (or readily adapt) the content to suit their own teaching emphases. Most teachers interviewed said that these types of tests fit best with their instructional schedules and curricula. And, from their points of view, these are the most valid instruments of those listed for such routine tasks as grading, on-going planning of teaching, etc. The predominance of locally developed tests at the secondary level supports the notion that high school teachers have more control over classroom assessment than do elementary school teachers. But heavy use of locally developed tests in the high schools may also reflect that they have fewer suitable commercial testing materials available. Comprehensive curricular programs -- including texts with coordinated workbooks, tests, etc. -- are more widely available for teachers of the elementary grades.

Finally, note that the two types of testing most often generated by state policy -- minimum competency testing and state assessment -- consume on the average very small proportions of classroom testing time.

The figures in Table 3 are averaged across all teachers in the survey, including those in states without minimum competency testing requirements. Even where minimum competency tests (MCT) are required in the grades sampled, however, less than three percent of the testing time at the sampled elementary grade levels and two percent of the testing time in secondary grade and subjects sampled is taken up by these tests. Where MCT's are available, but not required, they absorb less than one percent of the total testing time in the grades and subjects surveyed.

The picture with regard to statewide assessment programs is similar. Such programs require no more than three percent of the

Table 3

Types of Test Used, As a Percentage of the Total Time Devoted to Testing

TYPE OF TEST	Elementary Teachers		10th Grade English Teachers	10th Grade Mathematics Teachers
	Reading	Math		
Tests which form part of a statewide assessment program	3	3	5	1
Required Minimum Competency Tests	1	2	1	1
Tests included with curriculum materials	28	35	8	17
Other commercially published tests	17	18	6	3
Locally developed and district adopted tests	13	8	5	2
School or teacher developed tests	37	35	74	76

total testing time at the elementary level (or about 45 minutes per year on the average for reading and mathematics combined). At the high school level, tenth grade English assessment programs typically take about 75 minutes annually and mathematics programs an average of 30 minutes per year.

How are Test Results Used?

Long lists of tests' purposes have been provided in almost every test and measurement text in education. Lists of such purposes usually include selection, placement, remediation, instructional improvement, teacher assessment, accountability, and so on. But to what extent do these ideals represent reality? The survey questionnaires sampled a variety of potential purposes and examined the extent to which the results of particular types of tests and other methods of assessment actually serve each. Principals responded about the use of test results for school-level decision-making and communication, while teachers reported on classroom uses. The findings are summarized in Tables 4 and 5.

Principals reported about the importance of test results in eight specific areas. (See Table 4.) Based on the survey findings, it appears that principals ground their actions in all eight areas upon a wide range of information sources. Although no one of these sources is of overpowering importance, teachers' opinions and recommendations clearly carry more weight than do test results for each of the eight tasks listed. It appears that the more formal (and usually required) measures--standardized tests, minimum competency tests, and tests tied to district continua of instructional objectives--make their greatest

Table 4

Importance of Test Results for School Decision-Making in Elementary and Secondary Schools Reported by Principals*

ELEMENTARY

Decision Area:	A	B	C	D**	E	F
Curriculum Evaluation	3.01 (.67)	2.91 (.75)	3.04 (.87)	2.99 (.07)	2.94 (.84)	3.27 (.64)
Student Class Assignments	2.50 (.81)	2.35 (.91)	2.46 (.99)	2.44 (.08)	2.93 (.79)	3.12 (.71)
Teacher Evaluation	1.70 (.76)	1.53 (.78)	1.80 (.93)	1.68 (.14)	2.12 (.97)	----
Allocating Funds	1.91 (.87)	1.89 (.90)	1.94 (1.01)	1.91 (.03)	----	3.08 (.71)
Student Promotion	2.65 (.81)	2.31 (.96)	2.38 (.94)	2.45 (.18)	3.05 (.70)	3.29 (.67)
Public Communication	2.77 (.90)	2.47 (.99)	2.34 (1.00)	2.52 (.22)	2.31 (1.05)	----
Communicating to Parents	2.91 (.60)	2.64 (.98)	2.67 (.95)	2.74 (.15)	3.43 (.55)	3.45 (.57)
Reporting to District	3.12 (.68)	2.78 (1.10)	2.74 (1.10)	2.88 (.21)	2.62 (.91)	----

SECONDARY

Decision Area:	A	B	C	D**	E	F
Curriculum Evaluation	2.83 (.67)	3.27 (.64)	2.95 (.82)	3.02 (.23)	2.76 (.75)	3.14 (.70)
Student Class Assignments	2.77 (.77)	2.98* (.87)	2.78 (.87)	2.84 (.12)	2.98 (.73)	2.99 (.79)
Teacher Evaluation	1.63 (.74)	1.77 (.71)	1.84 (.78)	1.75 (.11)	2.39 (.83)	----
Allocating Funds	1.73 (.81)	2.20 (1.13)	2.06 (1.08)	2.00 (.24)	----	3.34ng (.54)
Student Promotion	1.61 (.78)	2.58 (1.28)	2.05 (1.13)	2.08 (.49)	3.33 (.85)	3.46 (.75)
Public Communication	2.84 (.80)	2.92 (1.03)	2.30 (1.07)	2.69 (.34)	2.24 (1.05)	----
Communicating to Parents	2.91 (.58)	3.03 (1.00)	2.55 (.99)	2.83 (.25)	3.56 (.55)	3.38 (.76)
Reporting to District	3.10 (.64)	3.12 (.97)	2.92 (.95)	3.04 (.11)	2.53 (.88)	----

A = Standardized, norm-referenced test batteries
B = Minimum Competency Tests
C = District Objective-based or Continuum Tests
D = Average Required Tests (A,B,C)
E = Results of Teacher and Curriculum tests
F = Teacher Opinions/Recommendations

[4-point scale: 4 = Crucial Importance - 1 - Unimportant or not used]

* Numbers in parentheses are standard deviations.
** Numbers in parentheses are standard deviations of values in columns A, B and C.

contribution in three tasks: curriculum evaluation, communicating with parents, and reporting to school-district personnel. Conversely, these types of tests are least important for teacher evaluation and in budget allocation. At the secondary school level, these more formal types of assessment (particularly the minimum competency tests) also play an important role in decisions about student class assignments. Further, while standardized, norm-referenced tests seem to be the most influential of the formal, required tests for principals at the elementary school level, minimum-competencytest results have more significance for high school principals.

Teachers also were asked to rate the importance of a variety of assessment types for activities in which they routinely engage. But while principals reported on assessment uses for school-wide activities, teachers were asked about assessment uses in four classroom tasks. (See Table 5.)

The results in Table 5 show that both elementary and secondary teachers do see test results of various types as useful in making a variety of decisions. Clearly, however, teachers accord the highest importance to their own observations of students' work and to their own clinical judgments. For initially grouping or placing students in a curriculum, for changing students from one group or curriculum to another, and for assigning grades, nearly every teacher respondent reported that their "own observations and students' classwork" is a crucial or important source of information. The great majority of- respondents also indicate that the results of the tests they themselves develop also figure as crucial or important in these decisions. Many

Table 5

Importance of Test Results for Teacher Decision-Making in Elementary and Secondary Schools*

Decision Area:	Standardized Test Batteries	District Continuim or Minimum Competency Tests	Tests Included with Curriculum	Teacher-Made Tests	Teacher Observations/ Opinions
			ELEMENTARY		
Planning teaching at beginning of the school year	2.53 (0.74)	2.60 (0.79)	----	----	3.39 (0.76)
Initial grouping or Placement of students	2.51 (0.74)	2.59 (0.82)	2.91 (0.74)	3.12 (0.83)	3.58 (0.78)
Changing a student from one group or curriculum to another, providing remedial or accelerated work	2.52 (0.79)	2.52 (0.81)	3.04 (0.74)	3.12 (0.84)	3.66 (0.72)
Deciding on report card grades	1.62 (0.76)	1.81 (0.81)	2.89 (0.79)	3.38 (0.74)	3.69 (0.72)
			SECONDARY		
Planning teaching at the beginning of the school year	2.22 (0.84)	2.38 (0.93)	----	----	3.59 (0.60)
Initial grouping or placement of students	2.28 (0.92)	2.46 (0.98)	2.48 (0.92)	3.04 (0.87)	3.84 (0.85)
Changing students from one group or curriculum to another, providing remedial or accelerated work	2.52 (0.95)	2.59 (0.86)	2.67 (0.93)	3.27 (0.76)	3.61 (0.66)
Deciding on report card grades	1.36 (0.66)	1.45 (0.64)	2.29 (0.96)	3.65 (0.62)	3.68 (0.65)

* [4-point scale: 4 = Crucial Importance - 1 = Unimportant or not used]

elementary school teachers also responded that the "results of tests included with the curriculum being used" are quite influential in their instructional decision-making.

Mirroring findings for principals, these results indicate that while teachers do not attribute heavy importance to the results of required tests, they do view them as somewhat useful sources of data for decisions about initial planning and placement of students in groups or curriculum, and even for decisions about reassigning students to different instructional groups or curricula throughout the year. In this last process, they probably serve as a kind of benchmark for judging individual student's "capabilities." For example, imagine a situation where a student is performing poorly in his or her instructional group. A teacher might examine standardized test results to determine whether the problem is "low ability" or whether other factors such as motivation seem a more likely explanation, and then base instructional decisions accordingly.

It is apparent from these results that teachers use a variety of sources to make each kind of decisions listed; they do not rely only upon a single information source. As one teacher stated:

> "You can't count a score on one test too heavily. The kid could be sick or tired or just not feel up to doing it that day. Maybe his parents had a fight the night before. Maybe he doesn't try. Maybe he doesn't test well." (Choppin, et. al. 1981)

Not only do survey respondents indicate that they consult several sources of information about students' achievement in making particular instructional decisions, respondents -- and particularly those at the

Table 6

Proportion of Teachers who Report Considering Many Types of Assessment Information Critical/Important for Given Activities

	Planning Teaching at Beginning of School Year	Initial Grouping or Placement of Students	Changing Grouping or Placement	Deciding on Report Card Grades
Number of Sources of Information Given in Question on Survey	4	7	6	6
Number of Sources Defined as "Many" for Purposes of this Analysis	3	4	4	4
Proportion of Elementary Teachers who Indicated That at Least this many functioned as Critical and/or Important for the Given Activity	50%	71%	62%	40%
Proportion of High School Teachers	33%	47%	49%	20%

elementary school level -- also report thinking that many kinds of assessment techniques give them crucial and/or important information. The data in Table 6 are illuminating here: over half the elementary school teachers surveyed report giving heavy weight to each of many sources of information in planning their teaching, in making initial groupings and placements, and in modifying instruction throughout the year.

What are Schools' and Districts' Administrative Practices in the Area of Testing and Test Use?

A growing literature suggests that district and/or school leadership is a significant determinant of whether and how educational innovations and practices are sustained (Berman & McLaughlin, 1978; Bank & Williams, 1982, Edmonds, 1979). Thus, the Test Use in Schools survey examined the practices of school and district administrators in: (1) making, and holding teachers accountable for curricular decisions based on test scores; (2) monitoring and/or supporting school and classroom testing practices; and, (3) providing information and staff development on testing. Exploratory fieldwork directed survey inquiry in these three general categories and (as was the case with other survey questions and item-response choices) suggested the particular items that were included in the instrument.

Making and holding teachers accountable for test-score-based curricular decisions. The school and district administrative practices in this area that were included on the survey appear in Table 7. Principals' and (where appropriate) teachers' responses regarding the frequency of each are reported in mean ratings on a four-point scale.* As the table shows, school and district administrators hardly ever establish specific test-score goals for individual schools or teachers. However, district administrators occasionally do check to see that areas in the curriculum that test scores indicate need improvement are in fact being emphasized in their schools; principals monitor their staff members' teaching fairly often toward this same end. Often, too (but not, on the whole, as a matter of routine), school administrators meet with teachers in groups or individually to review test scores and highlight their implications for curricular emphases.

It is worth noting that on the average, teachers report each of these practices as happening less frequently than principals do. It may be that the activities in question are more salient from principals' perspectives and less so from teachers'. Alternatively, principals may perceive them as more desirable practices than teachers do; if so, this perception may have led some principals to exaggerate the frequency of their occurrence.

Table 7 also indicates that test scores function in making and holding teachers accountable for decisions on curricular emphases less frequently at the secondary-school level than they do in elementary schools. Perhaps this occurs in relation to districts' practices in returning test results. Secondary principals find that scores are only rarely returned by their district such that they can be used in curricular decision making. In elementary schools, the curriculum-embedded tests that accompany basal reading and math series can be used as a basis for cross-classroom analysis of achievement patterns when standardized-test results and other scores are not forthcoming

---

* Mean rating on four-point scale: 4 = happens regularly, routinely; 3 = not regular or routine but happens fairly often; 2 = not regular or routine; it happens rarely; 1 = does not happen at all.

Table 7

Making and Holding Teachers Accountable for Test-score-Based Curricular Decisions

	Principals' Reports*		Teachers' Reports*	
	Elementary	Secondary	Elementary	Secondary
SCHOOL ADMINISTRATOR(S) . . .				
Meets with teachers to review scores and identifies areas that need extra emphasis	3.09	2.94	2.84	2.05
Observes teachers, reviews their plans to ensure areas indicated by tests are being emphasized	3.23	3.07	2.66	2.31
Takes test scores into account in evaluating teachers and/or establishes test-score goals for teachers to meet	1.57	1.55	1.46	1.27
DISTRICT ADMINISTRATOR(S) . . .				
Returns test results such that they can be used in school's curricular decision making	2.63	2.03	Not Asked	
Observes, reviews school plans and/or requires reports to assure school is emphasizing skills that test scores show need work	2.84	2.67	"	
Establishes specific test-score goals for school	2.12	2.33	"	

---

*Mean ratings on four-point scale: 4 = happens regularly, routinely; 3 = not regular or routine but happens fairly often; 2 = not regular or routine and happens rarely; 1 = does not happen at all.

from the district office. (Recall that the use of commercial, curriculum-embedded tests is more prevalent in the elementary grades.)

Monitoring and supporting testing practices. Table 8 displays those school and district practices examined in this area. Results are again shown as means on the four-point frequency scale. Of all the practices examined, only one seems to occur more than occasionally: district monitoring of the district testing program. Release time for teachers to develop tests is on the whole a rare phenomenon. So, too, are administrative reviews of (a) teacher-constructed tests and (b) student performance on such instruments as unit and chapter tests. (Although not specified in Table 8, the latter test types were mentioned explicitly in the questionnaire item.) These results suggest that there is little monitoring of teachers' classroom testing schedules. They also indicate that one type of measure upon which teachers rely heavily -- tests that they themselves construct -- is most often written individually and with no supervisory review.

Providing staff development and information about testing and test results. Principals were asked to comment on the frequency with which they and district administrators provided in-service experiences germane to testing and test results. In addition, teachers were asked to report on the occurrence of particular types of staff development over the last two years. The responses of principals and teachers to these questions are shown in Tables 9 and 10.

Table 8

Monitoring and Supporting Testing Practices

	Principals' Reports*		Teachers' Reports*	
	Elementary	Secondary	Elementary	Secondary
SCHOOL ADMINISTRATOR(S) . . .				
Requires teachers to turn in test scores/ grades on classroom tests and/or assignments	2.30 (1.10)	2.32 (1.10)	1.78 (1.17)	2.43 (1.02)
Requires teachers to turn in copies of tests they construct	1.62 (0.92)	2.17 (1.07)	Not Asked	
DISTRICT ADMINISTRATOR(S) . . .				
Conducts observations and/or requires reports to see that all aspects of district testing program are properly carried out	3.09 (0.95)	2.85 (1.07)	Not Asked	
Provides release time and/or extra pay for teachers to develop tests or curricular materials including tests	2.12 (1.03)	2.33 (0.98)	"	

---

*Mean ratings on four-point scale: 4 = happens regularly, routinely; 3 = not regular or routine but happens fairly often; 2 = not regular or routine and happens rarely; 1 = does not happen at all.

Table 9

Providing Staff Development and Information About Testing

Principals' Reports on Frequency*	Elementary	Secondary
SCHOOL ADMINISTRATOR(S) . . .		
Brings in speakers, workshops, printed material to update teachers' assessment skills	2.62 (0.87)**	2.48 (0.77)
DISTRICT ADMINISTRATOR(S) . . .		
Brings in speakers, workshops, printed material to update teachers' assessment skills	2.73 (0.98)	2.71 (0.90)

* Mean ratings on four-point scale: 4 = happens regularly, routinely; 3 = not regular or routine but happens fairly often; 2 = not regular or routine and happens rarely; 1 = does not happen at all.

** Numbers in parentheses are standard deviations.

Table 10

Percentages of Teachers Reporting Participation in Staff Development

Topic	Elementary	Secondary English	Secondary Math
(1) Analysis and explanation of state, district, or school test results	84	70	60
(2) How to administer tests required by my state, district, and/or school (procedures to follow, etc.)	78	54	46
(3) How to interpret and use results of different types of tests (e.g., norm-referenced and criterion-referenced tests and their applications)	59	35	34
(4) Alternative ways (other than tests) to assess student achievement	54	25	21
(5) How to tie what is taught more closely to the skills, content covered on required tests	50	37	25
(6) Presentation of published materials designed to prepare students for particular tests or to improve test-taking skills	41	32	29
(7) Training in the use of test results to improve instruction	35	21	19
(8) How to construct or select good tests	20	23	18

According to principals, staff development for teachers in the area of assessment occurs occasionally, i.e., with a frequency that on the average falls about midway between survey categories "very often" and "rarely." It appears that such staff development is generally initiated slightly more frequently by district administration than by principals.

Of all the topics listed, more teachers report participating in sessions devoted to: (a) analysis and explanation of test results, (b) directions for administering required tests, and (c) how to interpret and use the results of different types of tests. Staff development devoted to increasing teachers' routine classroom assessment skills, these data indicate, occurs much less frequently. Thus, for example, only about a fifth of the teachers in each category report receiving instruction in "how to construct or select good tests." Information on other means of assessment (alternatives to testing) was equally rare for secondary teachers, although some 54% of the elementary teachers did report staff development on this topic. Training in the use of test results to improve instruction was evidently provided for 35% of the elementary teachers and about 20% of the secondary teachers sampled.

Two other staff development activities on the list can be construed as aimed directly at improving students' test results. (See items five and six.) Between a quarter to a third of the secondary teachers have received training in these areas, while 40% to 50% of the elementary teachers have.

Finally, it is worth noting that secondary teachers, overall, report receiving staff development in topics related to testing less often than elementary teachers do.

Resources in support of testing. In a set of questionnaire items separate from those discussed just above, teachers were asked to comment on the availability and use of four resources which could support their classroom testing efforts. Teachers' responses to these items (Table 11) are presented in this section since the availability of each of these resources can be interpreted as due, at least in part, to the initiatives of school or district administrators. This is particularly true for item banks of test questions and computerized scoring and analysis of tests. In the case of the other two items included (other teachers with whom I plan and develop tests, someone to help grade tests and assignments), administrators can structure organizational arrangements that facilitate their availability and use.

The list of resources included in the survey instrument was selected on the basis of considerable fieldwork and piloting. Nevertheless, each resource was unavailable to a large proportion of respondents. The exception, of course, was "other teachers with whom I plan and develop tests or other evaluation assignments," but only about a quarter of the elementary-school teachers and a similar fraction of the secondary-school teachers reported taking advantage of this resource frequently. Some 45% of the secondary teachers reported constructing tests with others a few times a year, and fieldwork suggests that this often occurs as teachers in the same department conjointly devise mid-term and final exams.

Table 11
Available Resources for Testing Percentages of Teachers Reporting

Resource	NOT AVAILABLE	AVAILABLE Not Used	AVAILABLE Used Once To Several Times/Year	AVAILABLE Used at Least Once/Month	
Item banks of test questions upon which I draw in	71	4	8	16	Elementary
making up my tests.	51	8	24	16	Secondary
Other teachers with whom I plan and develop tests or other	37	12	26	24	Elementary
evaluation assignments.	21	10	45	24	Secondary
Someone who helps me read, grade, or correct	69	6	4	21	Elementary
tests and assignments.	70	5	4	21	Secondary
Quick, computerized scoring and analysis	64	2	30	4	Elementary
of tests	58	16	22	4	Secondary

Computerized test scoring and analysis was reported as used a few times annually by a quarter to a third of both the elementary and secondary teachers sampled. Fieldwork indicates that these reports may reflect the use of optical scanning machines for certain standard (including norm-referenced, standardized) tests. Some districts, however, have developed computer programs for scoring unit and chapter tests and simultaneously analyzing individual students' strengths and weakness on the skills they cover.

A final point: in general, nearly all those teachers who have access to the resources listed report using them at least sometime during the school year.

Teachers' and Principals' Perceptions of Tests and Testing

Fifteen declarative statements were included on the questionnaire to elicit teachers' and principals' perceptions of the quality, usefulness, and impacts of tests and testing. Respondents were asked to react to each statement by indicating whether they strongly agreed, agreed, disagreed, or strongly disagreed. Three of these items are irrelevant in this context. The proportion of teachers who strongly agreed or agreed with the remaining twelve statements is given in Table 12. Some of the findings presented in this table are highlighted below.

Perceptions of test quality. Survey responses suggest that teachers' opinions on the quality of commercial and district-developed measures are rather divided. A majority 59% and 62% respectively for the two types of tests) of elementary teachers indicated that they find both types of tests very sound--but a sizable minority (41% and 38% respectively) reached the opposite conclusion. Slightly less than half

Table 12

Teachers' Perceptions of Tests and Testing
Percentage of Teachers in Agreement With Each Statement

	TEACHERS		
QUALITY OF TESTS	Elementary	Secondary English	Secondary Math
Commercial tests are usually of high quality	59	46	46
The tests developed in our district are very good	62	62	60
The content (or skills) on most required tests is very similar to the content or skills that I teach	77	77	79
Tests of minimum competency are frequently unfair to particular students	58	48	35
USEFULNESS OF TESTING			
Testing motivates my students to study harder	73	80	93
Testing of minimum competency/proficiency/functional literacy should be required for promotion at certain grade levels or for high school graduation	81	86	90
IMPACTS OF TESTING			
Recently, I have been spending more teaching time preparing my students to take required tests	46	40	30
Tests of minimum competency have affected (would affect the amount of time I can spend teaching subjects or skills that the tests do not cover	62	62	42
In our school, testing programs are generally held to be much less important than the social problems with which we are concerned	39	32	42
As a result of minimum competency tests (and similar programs) parents are contacting schools about their children more frequently or in greater numbers	53	42	36
The pressure that testing exerts on the schools has a generally benefical effect	48	60	72
Teachers should not be held accountable for students' scores on standardized tests or tests of minimum competency	71	61	61

of the secondary teachers (46%) were convinced of commercial tests' quality, but a 60% majority supported the view that their district-developed tests are "very good."

It is impossible to know, of course, what criteria survey respondents use in judging whether or not these tests are of "high quality" or "very good," but other phases of Project inquiry provide some clues. Results of an earlier CSE questionnaire study of testing in five California school districts (Yeh, 1978) were reanalyzed in planning for the national survey under discussion here. Among the 256 elementary school teachers who responded on Yeh's instrument, the following appeared to be (in descending order) the most important criteria in test selection: similarity of test material to what was presented in class; clarity of test format; ease with which the test can be administered and/or scored. Fieldwork interviewees (Choppin, et. al., 1981), who also spoke of these considerations, emphasized too that they seek tests which yield information that they consider useful in their routine teaching tasks. The following quotations are illustrative.

> That computer-processed data [on district objectives-based tests] can really be used with those kids that need help. It does a better job [than the other tests available] of identifying students and students' needs...I can now say, "the kid needs to work on objectives 2, 3, 5, and 9."
>
> I don't feel we need to test, test, test; but if the information is something I can use to prescribe instruction, then I don't really mind giving it.

These and similar findings suggest that in judging the quality of tests, practical concerns (as opposed to technical, psychometric considerations) are foremost in teachers' minds.

Three quarters of the teachers in each survey category agreed that most "required tests" cover what they teach. This is one of the rare survey findings that is strikingly different than fieldwork results. Interviews both before and after the survey found many teachers complaining about the "mis-fit" between what they taught and material covered on standardized tests (which are usually required). Fewer interview respondents, but still more than the survey would suggest, commented on format and content differences between their texts and assessment instruments required by their state and district. It is possible, then, to speculate that survey respondents equated the term "required tests" with those that they themselves require of students (as many interviewees initially did), rather than with tests mandated by their district or state, as the survey intended. It is also possible (but we believe less likely) that our interviews were conducted in districts where teachers were unusually critical or that our interview questions inadvertantly "cued" a high proportion of negative reactions toward state and district tests.

Note that elementary-school teachers and teachers of high school English were more frequently critical of the fairness of minimum competency tests (MCTs). Issues of language and culture, among others, may be more salient for these teachers than for those of high-school mathematics, who on the whole found the fairness of MCTs less problematic.

Usefulness of tests. The great majority of teachers (73% of the elementary, 80% of the secondary English, and 93% of the secondary math) sampled indicated that they believe testing motivates their students to study harder. Perhaps with this in mind, an even larger proportion (81% of elementary, 86% of secondary English, and 90% of secondary math) agreed that proficiency or minimum competency tests should be required for promotion at certain grades or for high-school graduation.

Impacts of testing. Our fieldwork suggested that the very presence of testing--especially testing required by agencies beyond the school--would influence teachers' reports of trends in instruction. As the items in Table 12 under the "impacts" heading indicate, this was often the case. A substantial minority of teachers (from 46% at the elementary level down to 30% for secondary math teachers) reported that they have found themselves spending more teaching time preparing students for required tests. A near majority of teachers in each survey category (ranging from 62% of the elementary teachers to 42% of the secondary math) felt that minimum competency testing focuses (and probably contracts) their classroom curriculum in the direction of tested skills. And while many teachers seem to feel obliged to emphasize the skills that certain required tests cover, a great majority (ranging from 71% for elementary to 61% for secondary) reject the notion that they should be held accountable for students' performance on standardized and minimum competency tests. (Recall that many teachers interviewed during fieldwork portions of the study commented on the inappropriateness of weighing one assessment measure "too heavily," citing variations in students' motivation and test-taking skills as a rationale for their argument.)

While some teachers are apparently wary of testing's influence on curriculum, instruction, and their own accountability, opinion on these issues clearly is divided among respondents. Furthermore, on the whole, the proportions of the teachers in each survey category that express these concerns are roughly equaled by the those that cite benefits of testing. Slightly over half of the elementary-school teachers and over a third of those in high schools agreed that contacts with parents have increased as a result of minimum competency/proficiency testing programs. (Alerting parents whose children are in educational trouble is a typical feature of most MCT programs.) Nearly half of the elementary-school teachers (48%) and a substantial majority of the high-school teachers (60% of English and 72% of math teachers) also concurred with the proposition that "the pressure that testing exerts on the schools has a generally beneficial effect."

We began this paper by citing the controversy over achievement testing that has arisen in academic circles through the last six or eight years. The results reported in Table 12 suggest that present achievement tests and testing practices may be equally controversial among educators in the schools. At the very least, the perceptions of the teachers are mixed with respect to the quality and impacts of tests and testing. It may be that the perceptions of individual teachers are finely differentiated and highly complex, reflecting considerable thought. Alternatively, the patterns of response to these questions may signify that many teachers currently hold ambivalent, or even contradictory, viewpoints with respect to the merits of testing.

Principals' perceptions of testing and test use. A brief discussion of principals' views will complement the foregoing discussion. Principals responded to a set of statements which included some of those presented to teachers and some designed exclusively for administrators.

Most principals seem to be satisfied with the quality of available tests: over 80% agree that "standardized tests are fair for most students" and that the quality of both district developed tests and commercial curriculum tests is generally good. Almost half, however, express concern about the equity of minimum competency tests for some students, and a sizeable minority (43%) have reservations about the "pressure that required testing exerts upon me and the teachers in my school." Nonetheless, most feel that "test scores are a fairly good index of how well a school is doing," (64%) and that schools should be held accountable for their students' scores on standardized achievement tests (60%) and on minimum competency tests (73%). They are on the whole uncomfortable with the idea of using test scores to evaluate teachers: over 60 % of the elementary school principals and a bare majority of secondary principals agreed that test scores should not be "used to evaluate teachers' effectiveness or competence."

A majority of the principals surveyed report satisfaction with the amount of time devoted in their schools to "required testing and the preparation for it." More than half advocate required minimum competency testing for grade promotion and high school graduation.

What Factors Influence Testing Practices?

The findings presented thus far have been descriptive of national values for elementary and secondary teachers and principals. As indicated previously these values are the result of weighted computations designed to estimate the actual numerical values for the respective populations of interest (elementary teacher or principal, secondary teacher or principal). While providing these point estimates of national test use results was one of the primary objectives of the Test Use Study, another objective was to explore and identify relationships that impinge on test use in the schools. That is, we were concerned with investigating the relationships between test use and certain policy relevant variables. In so doing, it was hoped that a framework could be developed that would both integrate the results of the current study as well as guide future studies of this topic.

As the nature of this effort was exploratory and interest was in identifying relationships rather than projecting specific values, it was decided that unweighted analyses should be performed. Thus, the results reported in this section should not be construed as actual projections of national values. Rather, the results should be interpreted as indicative of likely relationships that may exist in the schools nationwide.

The exploratory analyses were conducted in two phases. In the initial phase, we examined the relationship between three key policy variables (district minimum competency testing requirements, district socio-economic status, and school context) and a variety of test-use indicators developed from the survey results including amount of testing, use of test results, and perceptions of testing.

Analyses utilized scales created to examine various aspects of achievement testing practice including:

- Amount of total student time on testing (in minutes) as reported by teachers.
- Use of assessment results as reported by teachers: i.e., the importance attributed to results summed over all decision areas.
  - ° use of formal measures, including norm-referenced, standardized tests, minimum competency tests, and district-objectives-based tests
  - ° use of curriculum-embedded testing, including placement, chapter or unit, and end-of-book or end-of-level tests
  - ° use of teacher-made tests
  - ° use of teacher judgment
- Perceptions of testing as reported by teachers.*
  - ° quality and value of tests
  - ° equity and desirability of minimum competency tests
  - ° emphasis on basic skills (as it co-occurs with different testing practices and other variables).

For each policy variable a series of analyses were performed investigating the relationship of that variable to each of the survey indicators.

---

* Composite variables were created to represent the three general subcategories included in the fifteen perception-elicitation statements discussed earlier. Thus, the quality/value composite was based upon respondents' mean rating (on the four-point scale where 4 = strongly agree; 1 = strongly disagree) across six perception items; the MCT equity/desirability scale on mean responses to two items; and the basic-skills emphasis scale on mean responses across four items.

In the second phase, we developed a conceptual model of test use in the schools which integrates the results of the first phase. Further, this model was subjected to a formal analysis in order to determine its reasonableness. In the pages below the results of the first phase analyses for the three policy variables are presented prior to the examination of the conceptual model tested in the second phase.

The relationship of minimum competency testing requirements and testing. It appears that minimum competency testing requirement may have affected the time spent on testing. Controlling for other sampling factors, there is less testing where minimum competency tests are required. This difference is dramatic at the secondary level where all types of minimum competency requirements seem to be accompanied by less testing in the schools and where requirements for promotion or graduation are accompanied by the lowest amounts of testing time. (See Table 13.)

Teachers in environments where minimum competency tests are required for promotion/graduation do not have a significantly different view of tests and testing in general than those in other settings. Their attitudes on the quality of tests and value of testing are neither more positive nor more negative than those where MCTs are required only for diagnosis or in settings where MCTs were not required. However, their views toward minimum competency testing do differ. Teachers where MCTs are in fact required for promotion or graduation are more circumspect about the equity of such tests for all students and about the advisability of requiring MCTs for promotion. These findings may signal problems that appear to teachers when such requirements become a

Table 13

Relationships between Minimum Competency Testing Requirements and Total Time in Testing

Reported in Minutes

	SECONDARY			ELEMENTARY		
	English	Math	Total per Teacher[1]	English	Math	Total Per Teacher
No Mimimum Competency Testing (MTC)	3723.53	3173.38	3455.01	577.45	570.91	1148.37
MCT required for diagnosis, state-mandated measure	915.77	1180.50	1086.47	504.32	448.15	992.48
MCT required for diagnosis, local choice of measure	1600.07	1394.57	1482.77	489.90	486.32	976.22
MCT required for promotion or graduation, state measure	1427.73	808.15	1095.86	338.69	632.88	971.57
MCT required for promotion or graduation, local choice of measure	766.78	786.29	769.87	401.98	625.85	1027.84

[1] Difference in mean values of different MCT categories statistically significant at $p < .01$

reality. (See Table 14.) Note too, that as consequences grow more serious, i.e., elementary promotion vs. secondary graduation, teachers' views apparently grow more cautious. Perhaps as a result of these consequences, secondary teachers where MCT's are required for promotion or graduation find a greater emphasis on basic skills instruction and a greater need to emphasize tested skills than do other teachers in the sample. These trends were not observed at the elementary school level.

Table 14

Relationships between Minimum Competency Testing Requirements and Attitudes Toward Minimum Competency Testing*

	SECONDARY 1	ELEMENTARY 2
MCT required for promotion/graduation, state-mandated measure	3.56	4.24
MCT required for promotion/graduation, local measure	3.76	4.29
MCT required for diagnosis, state measure	3.93	4.38
MCT required for diagnosis, local measure	4.20	4.96
No MCT	4.16	4.79

1 $p < .05$
2 $p < .01$

* Values on this scale ranged from 2 to 8, with a value of '2' indicating a strong negative attitude and a value of '8 indicating a strong positive attitude.

While differences related to minimum-competency-testing status were observed in amount of time spent testing and in attitudes toward tests, no differences were found in the use of test results. That is, despite the consequences of minimum-competency-testing programs, teachers do not report according more importance to test results in general. This may suggest that minimum competency efforts are separate from mainstream instruction.

The relationship of socio-economic status to testing. Given the evaluation and testing requirements associated with compensatory programs, it seemed likely that students from low SES backgrounds would be subjected to more testing -- and therefore lose more instructional time -- than their more advantaged peers. However, available data indicate that students in lower SES areas do not spend more total time in testing than those in middle- and upper-income settings, nor do they spend more time in required testing. In fact, there is no relationship between total test time and SES when either a district or a school level indicator is employed.

Teachers' use of test results also appears unrelated to the socio-economic status but differences do occur in principals' reported uses Test results apparently have greater impact and wider consequences in lower SES schools than they do in higher SES settings. In the latter, principals report paying more attention to test scores, particularly those of minimum-competency and district-continuum tests, in evaluating curriculum, deciding on student class assignments, allocating funds, and in communicating with and reporting to the public, parents, and the district. (See Table 15.)

Table 15

Importance of Test Results for School Decision-Making in Schools of Higher and Lower SES*

Decision Area:	Standardized norm-referenced test batteries	Minimum Competency Tests	District Objective based or Continuum Tests	Average Required Tests (A,B,C)**
HIGHER SES				
Curriculum Evaluation	2.90 (.52)	2.95 (.71)	2.64 (.92)	2.83
Student Class Assignments	2.49 (.71)	2.24 (.79)	2.10 (.96)	2.27
Teacher Evaluation	1.69 (.72)	1.81 (.74)	1.94 (.81)	1.81
Allocating Funds	1.85 (.83)	1.85 (.91)	1.71 (.86)	1.80
Student Promotion	2.19 (.83)	2.49 (1.04)	2.27 (.95)	2.31
Public Communication	2.69 (.78)	2.36 (.96)	2.33 (1.00)	2.46
Communicating to Parents	2.80 (.56)	2.74 (.94)	2.51 (.84)	2.68
Reporting to District	3.03 (.73)	2.94 (1.09)	2.74 (.94)	2.90
LOWER SES				
Curriculum Evaluation	3.08 (.78)	3.18 (.59)	3.08 (.83)	3.11
Student Class Assignments	2.68 (.79)	2.67 (1.03)	2.59 (.94)	2.65
Teacher Evaluation	1.95 (.84)	1.74 (.72)	1.94 (1.03)	1.88
Allocating Funds	2.00 (.79)	2.45 (.92)	2.18 (1.00)	2.21*
Student Promotion	2.45 (.93)	2.39 (.99)	2.17 (.84)	2.34
Public Communication	2.84 (.90)	2.93 (.97)	2.59 (1.04)	2.79
Communicating to Parents	2.96 (.57)	3.26 (.78)	3.26 (.51)	3.16
Reporting to District	3.11 (.65)	3.28 (.61)	3.11 (.93)	3.17

[4-point scale: 4 = Crucial Importance - 1 = Unimportant or not used]

* Numbers in parentheses represent standard deviations.
** Numbers in parentheses represent standard deviation of values in columns A, B and C.

Consistent with trends in the use of test results at the school level, testing receives greater administrative attention in lower SES schools than in those serving more affluent communities. Particularly noteworthy is the emphasis and support principals perceive at the district level -- in terms of both oversight and useful reporting practices. Districts apparently are also far more likely to establish specific test-score goals for lower SES schools. Principals in 40% of these schools report that their districts routinely set such goals, while only 10% of principals in higher SES schools so report.

This emphasis on test scores at lower SES schools is also accompanied by a greater curricula emphasis upon basic skills. While a simple majority of principals feel that more time is being spent in basic-skills instruction (reading English, and math) as a result of testing programs, over 80% from lower SES schools perceive this increased emphasis. Teachers in lower SES environments likewise find a greater emphasis on basic skills instruction and more need to emphasize tested skills than do those in higher SES communities. It appears that tests, through a perception of a basic skills press, may be driving the curriculum in low SES schools.

The relationship of school contextual factors to testing. The study examined three factors in the local school environment which, on the basis of earlier research, seemed likely to influence teachers' testing practices: staff development; local leadership practices; and availability of resources for testing and instructional differentiation. Scales representing each of these variables were developed by aggregating related questionnaire responses, and their relationships to amount of, use of, and attitudes toward testing were examined. Correlational analyses indicate that all three factors are significantly related to some aspect of teachers' testing practices, though none were related to the amount of time spent on testing.

The information and training about tests factor reflects how much information and training through staff development activities, teachers received in the last two years. It was hypothesized that knowledge about testing and how test results can be utilized in the classroom setting could facilitate teachers' use of tests and/or influence their attitudes toward testing. The correlative analyses support these hypotheses, particularly at the elementary-school level. More training is associated with greater use of formal tests for instructional decision-making and with more positive attitudes towards the quality and utility of tests. (See Table 16.) Amount and diversity of staff development, however, are not related to the use of curriculum-embedded or teacher-made tests-- probably because the kinds of inservice training teachers report usually focus on more formal measures.

Curricular accountability is likewise related to test use and attitudes. Survey results indicate that when principals show that they care about test scores -- by reviewing test scores to identify curricular weaknesses, taking action to assure teachers are emphasizing skills that test scores show are needed, etc. -- teachers pay more attention to tests in their instructional planning and feel more positively about the usefulness of tests.

Survey findings also indicate that testing resources such as someone to help correct or grade tests; quick, computerized test

Table 16

Relationships Between Contextual Factors and Testing Practices

	STAFF DEVELOPMENT				LEADERSHIP SUPPORT				INSTRUCTIONAL RESOURCES				TESTING RESOURCES			
	Elem.		Sec.		Elem.		Sec.		Elem.		Sec.		Elem.		Sec.	
	R	M	E	M	R	M	E	M	R	M	E	M	R	M	E	M
Attitude Toward Quality of Tests	.318	.206	.215	___*	.230	___	.206	___	___	___	___	___	___	___	___	___
Use of Formal Testing	.350	.300	.198	.256	.219	.235	.163	.333	.171	.288	.207	.230	.229	.340	.126	.220
Use of Continuum Embedded Tests	___	___	___	___	___	___	.156	.376	.254	.391	.215	.236	.232	.361	.286	.237
Use of Teacher Made Tests	___	___	___	___	___	___	___	___	.206	.430	___	___	.241	.362	___	.176

* Statistically non-significant (p. 2.05) correlations have been indicated with a ' _ '.

79

78

scoring and analyses; item banks of test questions; or collaborative arrangements for test development are not widely available. Nevertheless, the greater the number of these resources that are available, the greater the importance teachers accord to all kinds of assessment results, including their own observation-based judgments.

The use of test results for instructional planning and decision-making assumes that some action can be taken on the basis of student test scores -- e.g., providing remediation or advanced work for individual or small groups of students. Instructional resources, such as aides, instructional machines, and alternative curriculum materials must be available to make such actions feasible; where there are no options, no decisions are necessary and likewise test scores indicating the need for alternative actions are superfluous. Survey findings support this logic: availability of instructional resources is related to the use of all kinds of tests at the elementary school level and to the use of formal and curriculum embedded tests at the secondary level.

A Conceptual Model for Teacher Test Use

The previous section presented the results of a series of exploratory analyses designed to identify possible relationship between certain meaningful constructs and total test time and test use. These results indicated that no consistent pattern of relationships with total testing time were evident at either the elementary or secondary level. However, several relationships were found between the use of certain types of tests for instructional decision-making by teachers and some of the constructs from the previous section. This section examines these relationships within the framework of a single conceptual model that would capture the important policy implications of these associations. It should be stressed that while this examination was conducted using the techniques of path analysis, the results should not be construed as anything more than indicative. Because of the exploratory nature of the analyses no formal tests of the conceptual model or of alternative models were conducted; rather only single relationships (paths) were tested for statistical sisgnificance. Thus, while the model to be presented shows significant relationships between the constructs, it is not necessarily the only possible explanation for these relationships. The remainder of this section is organized by the results of the path analyses for elementary and secondary teachers.

Elementary Teacher Test Use

The conceptual model shown in Figures 1 and 2 (see Appendix) incorporates the results for four different outcomes reflecting teachers' use of different types of assessment. That is, relationships between the teacher use of specific test types and the *policy variables* were explored within the same model. As can be seen in those figures, four types of decision-making devices were included: formal standardized tests, curriculum embedded tests, teacher-made tests, and teacher observations/judgments. For each of these, we examined the relationships between amount of use and variables including: *perception of basic skill press*, attitudes about quality of tests, testing resources, instructional resources, information about tests, *curricular accountability*, and school level socioeconomic status. It was hypothesized the school SES would act as an exogenous variable in this system of relationships. Further, it was thought that *curricular* accountability on the part of the principal would drive the amount of information and training received by the teachers. That is, participants who were viewed as emphasizing and supporting greater use of tests were also likely to provide and require more training on test use. Lastly, it was assumed that accountability and information would relate to attitudes about test quality and basic skills press.

The tenability of these hypotheses can be ascertained from the results presented in Figures 1 and 2, displaying results of elementary school reading and mathematics. The paths drawn in these figures represent statistically significant regressions between the variables involved. Paths not drawn in the diagram indicate that the regression was not statistically significant.* Looking at the results in these two figures, one is struck by the high degree of correspondence. In fact, there is only one relationship that was statistically significant in one case and not the other. For elementary math teachers there is a significant relationship between the amount of instructional resources and use of formal tests in decision-making while that relationship does not appear for reading teachers. With that exception the two models are identical in their structure indicating that the same mechanism is likely to be operating regardless of subject matter.

Beyond the concordance between the two cases there are several interesting features of the model. First of all, the influence of SES on the use of tests in decision-making is moderated through variables which are directly under administrative control. Specifically, the amount of information and training about tests, and the degree to which the principal holds teachers accountable, moderate the influence of SES on test use. Thus, regardless of a school's SES it appears possible through administrative steps to influence a teacher's use of tests. This administrative effect appears to be manifested through the attitudes that teachers have about tests. In particular, teachers seem to have better attitudes about the quality of tests in schools where there is more information and training about tests. Additionally, teachers who are more informed about tests and are held more accountable by the principal for test results also perceive a greater emphasis on basic skills and basic skills tests. These characteristics translate into greater use of formal testing in making classroom decisions.

---

* A probability level of .05 was used in these analyses to determine statistical significance. The single exception to this criteria has been noted in the Figures. The basis for this exception was the exploratory nature of the analysis which generally involves somewhat more lenient criterial for examination of results.

The use of formal tests is also a function of the amount of resources available to the teacher. The greater amount of testing resources (e.g., scanning, scoring help) the greater the use of formal testing. Further, increased instructional resources leads to greater use of formal testing. The hypothesis here is that resources permit instructional alternatives or options. The existence of these options requires greater decision-making on the part of teachers and hence greater use of test results.

The use of curriculum embedded tests seems to be a function of the amount of both testing and instructional resources as well as the teacher's perception of the quality of tests. In situations where the teacher feels that the commercial tests are well made they will be more likely be employed in decision-making. Again, the role of resources seems to be one of making testing or test use more feasible.

It is interesting to see in the results of these analyses that the only contributing factors to the use of teacher-made tests and teacher judgment are the resources available to the teacher. This finding may reflect the pervasive use by teachers of these mechanisms for arriving at instructional decisions almost independent of other sources of information. That is, there may be a feeling on the part of teachers that their own tests and judgments are more suitable for decisions than more formal measures regardless of their attitudes and training about these latter tests.

In sum, the model portrayed in Figures 1 and 2 shows that the use of test information in teacher decision-making can be influenced by administrative action. In particular, the administrator can require greater accountability on the part of the teachers, provide more information and training about tests and, if feasible, supply additional testing and/or instructional resources. Each of these actions appears to positively influence the use of one or more types of test use.

Secondary Teacher Test Use

Similar analyses were performed for secondary school teachers who taught English (reading) and Mathematics. The results of these analyses are presented in Figures 3 and 4 in the Appendix. As can be seen from these figures the picture at the secondary level is not nearly as clear nor consistent. In fact, there are few statistically significant relationships for the English teachers and those that do exist are for the use of curriculum tests. Because of the paucity of relationships for these teachers it would be hazardous to attempt to interpret them or the model.

The results for mathematics teachers are somewhat more encouraging though still not as conceptually appealing as the elementary school results. The results in Figure 4 show that a somewhat similar mechanism to that found in elementary schools may be operating for the use of formal and curriculum tests. That is, it appears that curricular accountability, information about tests, and testing resources are all influencing the use of formal and curricular tests. What appears to be different at this level, however, is the greater direct role of curricular accountability. This variable has strong direct relationships to both use variables. Further, this variable, rather than information about tests, seems to relate to teachers' attitudes about test quality. Thus, these results seem to point to greater importance of the role of the principal in establishing curricular accountability than at the lower grade levels. It should be noted, however, that the same constraints are still involved with use of tests, it is just their relative priorities and interrelationships that are different. Therefore, from a prescriptive point of view, working on the three variables of information and training about tests, curricular accountability, and testing resources seem most likely to pay off in terms of greater teacher use of formal and commercial tests.

In summary, these analyses have explored a possible prescriptive model for teacher use of different types of information in their decision-making. While the results showed some disparity between elementary and secondary teachers, particularly for secondary English teachers, some definite similarities were found. In particular, it appears that three policy relevant and administratively manipulatable variables are related to increased use of formal and commercial tests. These three variables are the amount of curricular accountability operating in the school, the amount of information and training given to the teachers about tests, and the amount of testing related resources made available to the teacher. It would appear that if increased use of formal test results was a desirable goal, increased emphasis should be placed in the three areas mentioned above.

Concluding Remarks

As we conclude our analyses of Test Use in Schools Project data, we are left with the feeling that considerable additional information and investigation are needed to understand more fully and to model those factors that most influence local testing practices. However, the data from this study have identified many important areas which seem to influence testing. In particular, features of the school environment are among the most influential in determining how much attention teachers give to the results of formal testing. Further, project findings also suggest some of the qualities that teachers seek in tests -- qualities which local educational agencies might strive to embody in their testing programs. Other results indicate the advisability of attending to the quality and extent of pre-service and in-service teacher training in assessment. And still others add to our understanding of the ways in which teachers think and reason as they carry out routine classroom tasks.

The specific findings of this study are presented in summary and narrative form below:

CSE STUDY OF TEST USE

Summary of Findings

1. **How much basic skills testing goes on in schools?**

   a. The typical upper-elementary-grade school student spends about:

   - 10 hours a year in reading tests
   - 12 hours a year in mathematics tests
   - 5% of instructional time in testing in each subject

   b. The typical secondary student spends about:

   - 26 hours a year in English tests
   - 24 hours a year in mathematics tests
   - 20% of instructional time in testing in each subject

   c. Student test time represents only about 1/4 to 1/3 of the time teachers spend in tests related activities.

   d. Secondary students spend less time in testing where minimum competency testing is required for promotion. Student time in testing is unrelated to any other sampling factor, including SES.

2. **What kinds of basic skills tests are administered?**

   a. Elementary school teachers report:

   About half of the testing they conduct is required by their state or school district.

   Teacher-developed tests and commercial curriculum-embedded tests each account for about one-third of classroom testing.

   Required minimum competency tests account for a very small percentage (3-5%) of test administration time in the grades studied.

   b. High school teachers report:

   About one-quarter of the testing they administer is required by their state or school district.

   The majority of testing (75% in English and in mathematics) is teacher developed.

   Minimum competency testing accounts for only a small portion of the testing conducted.

3. How are test results used?

a. It is clear that principals and teachers base their actions and decisions upon a wide range of information sources. No one testing source is of overwhelming importance: greatest weight is accorded to professional observations and opinions.

b. Teachers and principals find test results as useful and at least moderately important in making a variety of decisions:

- Principals report that formal tests -- standardized tests, minimum competency tests and tests tied to district continua -- are most influential for three tasks: curriculum evaluation, communicating with parents, and reporting to school district personnel. These types of tests are little used for teacher evaluations or in budget allocation.
- Teachers report that the results of formal tests are moderately useful for planning teaching at the beginning of the school year, for initially grouping or placing students in a curriculum, and for changing a student from one group or curriculum to another and in identifying needs for accelerated or remedial work. Teacher developed tests and, at the elementary school level, curriculum-embedded tests, play a strong role in each of these decision areas as well as in deciding on grades.
- Secondary teachers accord less weight to formal and curriculum-embedded tests than do elementary school teachers.

c. Teachers and principals in lower SES schools seem to accord slightly more importance to test results than do those in higher SES schools.

4. What are schools' and districts' administrative practices in the area of testing?

a. Accountability in test-score based curricular decisions:

- While school and district administrators rarely (except for lower SES schools) establish specific test score goals for individual schools or teachers, they do check to see that areas in the curriculum which tests scores indicate need improvement are in fact being emphasized.
- School administrators likewise meet with teachers fairly often to review test scores and highlight their implications for curricular emphases.
- Secondary teachers are less accountable to test scores for curricular planning than are their elementary peers.

b. Monitoring and support of testing practice:

- There is little monitoring of teachers' classroom testing practices.
- Few resources -- e.g., release time to develop tests, aides to help grade tests, access to item banks, quick or computerized scoring -- are available to support testing activities.

c. Providing staff development and information about testing and test results:

- Secondary teachers receive less information and training related to testing than do elementary school teachers.
- For elementary school teachers:

  A great majority receive information or training in how to administer tests required by the state, district, and/or school and analysis and explanations of the results of such tests.

  About half receive information or training in how to interpret and use the results of different types of tests and in alternative ways to assess student achievement.

  About half receive information or training related to raising test scores: how to tie what is taught more closely to the skills covered on required tests and published materials designed to prepare students for particular tests or to improve test taking skills.

  Few receive training in how to construct or select good tests or in how to use test results to improve instruction.
- For secondary teachers:

  Most receive analyses and explanation of state, district ,or school test results, a bare majority receive information on how to administer required tests, and only a minority receive information or training in any of the other listed areas.

d. District and school administrative practice appear related to testing practices and attitudes toward testing.

5. What are teachers' and principals' attitudes toward the quality of tests?

a. Principals' and teachers' attitudes toward testing are divided. While a majority appear relatively pro-testing, a sizeable minority of teachers (sometimes approaching 50%) express serious reservations about required standardized and minimum competency tests.

b. Most teachers (60%) and the great majority of principals feel that tests developed by their district are very good, and similar proportions of principals and elementary teachers likewise agree that commercial tests are of high quality. Less than half the secondary teachers are convinced of the quality of commercial tests.

c. Three quarters of the teachers at each level agree that most "required tests" cover what they teach.

d. More staff development and greater administrative support for testing are associated with more positive attitudes toward the quality of tests--for both elementary and secondary teachers.

6. **What are teachers' and principals' attitudes toward minimum competency testing?**

a. A substantial proportion of principals as well as elementary and high school English teachers are critical of the fairness of minimum competency tests for some students--particularly in those schools where minimum competency tests are in fact required for promotion or graduation. Fewer high-school mathematics teachers (35%) express concern.

b. Most teachers agree that minimum competency tests should be required for promotion at certain grades or for high school graduation. Principals appear more circumspect about MCT: about half advocate such minimum competency testing requirements.

c. Teachers are less supportive of MCT as a requirement for promotion or graduation where MCT is currently required for these purposes.

d. Elementary teachers in lower SES schools hold less positive views about minimum competency testing than their peers in more advantaged settings.

7. **What are principals' and teachers' views about the impact of testing on the school curriculum?**

a. A sizeable minority of teachers (almost 50% at the elementary school level) note that they recently have been spending more teaching time preparing their students to take required testing.

b. About 60% of the sampled teachers (excluding secondary math teachers) assert that minimum competency testing affects the amount of time devoted to content or skills not covered by the tests.

c. Secondary teachers where minimum competency tests are required for promotion or graduation find a greater emphasis on basic skills testing and greater need to emphasize tested skills.

d. Principals, to a greater extent than teachers, believe that required testing programs result in more time being spent in basic skills instruction, particularly in lower SES schools where 80% so reported.

e. The impact of testing on the curriculum is negatively related to socio-economic status, with greater impact in lower SES settings.

8. What are principals' and teachers' views about testing and accountability?

a. A great majority feel that teachers should not be held accountable for or evaluated by their students' performance on standardized tests.

b. Most principals feel that test scores *are* a fairly good index of how well a school is doing and that schools should be held accountable for their students' test scores.

9. What factors influence the use of test results?

a. Teachers' use of formal test results is related to their attitudes toward tests, staff development training, and the testing and instructional resources available to them.

b. Teachers' use of curriculum embedded tests is related to their attitudes about test quality and the instructional and testing resources available to them.

c. At the elementary school level, teachers' use of teacher-developed tests and their own observations and judgments is related to available resources and staff development opportunities. Lower SES settings are associated with greater use of teacher-developed tests.

d. Socio-economic status seems to be related indirectly to use of test results, through its relationship with staff development, test-score based curricular accountability, and perceptions of basic skills curricular emphases. Lower SES settings are associated with more staff development, greater curricular accountability, and heightened perceptions of a basic skills press in the curriculum.

APPENDIX

Tables A1 - A2
Figures 1 - 4

Table A1

School Characteristics

	Elementary		Secondary	
	Mean	S.D.	Mean	S.D.
Total Enrollment	528	(235)	1439	(696.3)
School Ethnicity				
Black	15.0%	(25.8)	15.0%	(25.5)
Hispanic	8.1%	(21.2)	6.8%	(18.4)
Asian	2.1%	(9.2)	0.7%	(1.2)
Native American	5.5%	(20.4)	0.4%	(2.1)
Caucasian (Euro-American)	70.6%	(35.8)	76.2%	(31.0)
Other	1.2%	(9.9)	0.7%	(5.7)
Socio-Economic Status				
Low income (< $8,000)	32.2%	(26.2)	22.4%	(20.2)
Middle income	51.6%	(23.4)	56.7%	(19.3)
High income (> $25,000)	20.5%	(21.7)	21.8%	(17.6)
% of student receiving AFDC or free lunch	31.0%	(26.2)	23.2%	(22.8)
Transiency Rate	15.5%	(13.7)	10.4%	(7.8)
Absentee Rate	6.0%	(9.4)	7.4%	(3.7)
School Improvement Program				
% Participating	59.7%	---	63.0%	---
% Requiring Testing	76.3%	---	65.7%	---
Minimum Competency Testing				
Required	53.3%	---	50.0%	---
% Students passing first time	80.0%	(23.0)	76.1%	(22.6)

Table A-2

Teacher Characteristics

	Elementary	Secondary
Average Number of Years of Teaching Experience	12.03 (7.50)	2.69 (7.50)
Average Number of Years of Teaching in District	9.68 (6.94)	10.04 (7.00)
Percentage of Teachers whose Highest Diploma is:		
Bachelors	57.92	50.66
Masters	41.65	48.44
Doctorate	0.17	0.91
Average Numbers of credits/units beyond last degree	24.10 (24.39)	25.82 (22.34)
Average Number of students in class	27.11 (9.45)	26.09 (9.84)
Average Hours per week of English or Reading	6.55 (1.97)	5.38 (1.78)
Average Hours per week of Mathematics	5.19 (1.44)	5.62 (1.67)

ELEMENTARY READING

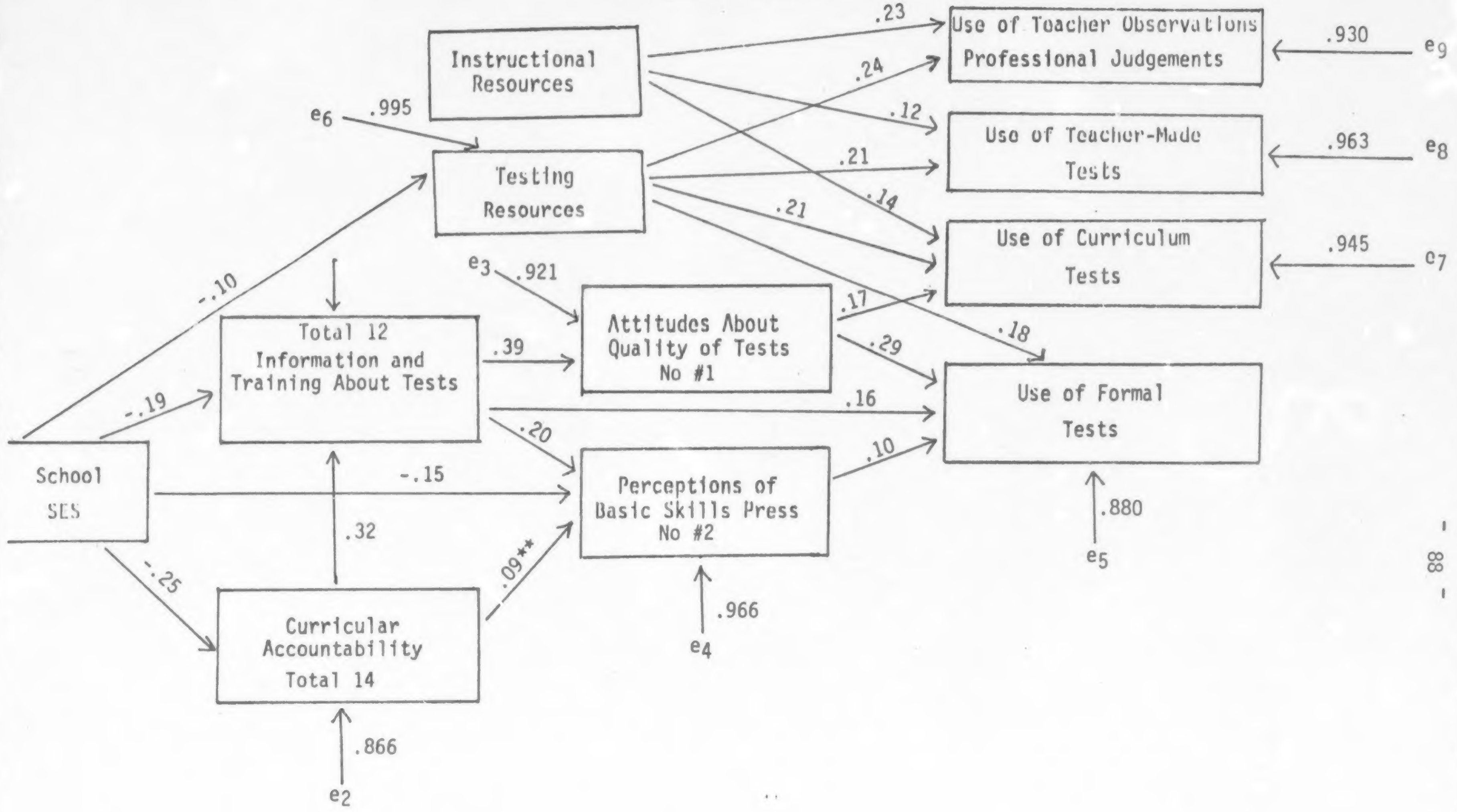

Instructional Resources
Testing Resources
e6
.995
Total 12 Information and Training About Tests
Attitudes About Quality of Tests No #1
e3
.921
Perceptions of Basic Skills Press No #2
e4
.966
Curricular Accountability Total 14
e2
.866
School SES
Use of Teacher Observations Professional Judgements
e9
.930
Use of Teacher-Made Tests
e8
.963
Use of Curriculum Tests
e7
.945
Use of Formal Tests
e5
.880
.23
.24
.12
.21
.14
.21
.17
.18
.29
.16
.10
.39
.20
-.10
-.19
-.15
-.25
.32
.09**

FIGURE 1

CONCEPTUAL MODEL FOR ELEMENTARY SCHOOL TEACHERS'TEST USE IN READING*

95 *Reported values correspond to standardized path coefficients that were statistically significant ($p < .05$)

**Reported coefficient statistically significant ($p < .06$).

ELEMENTARY MATHEMATICS

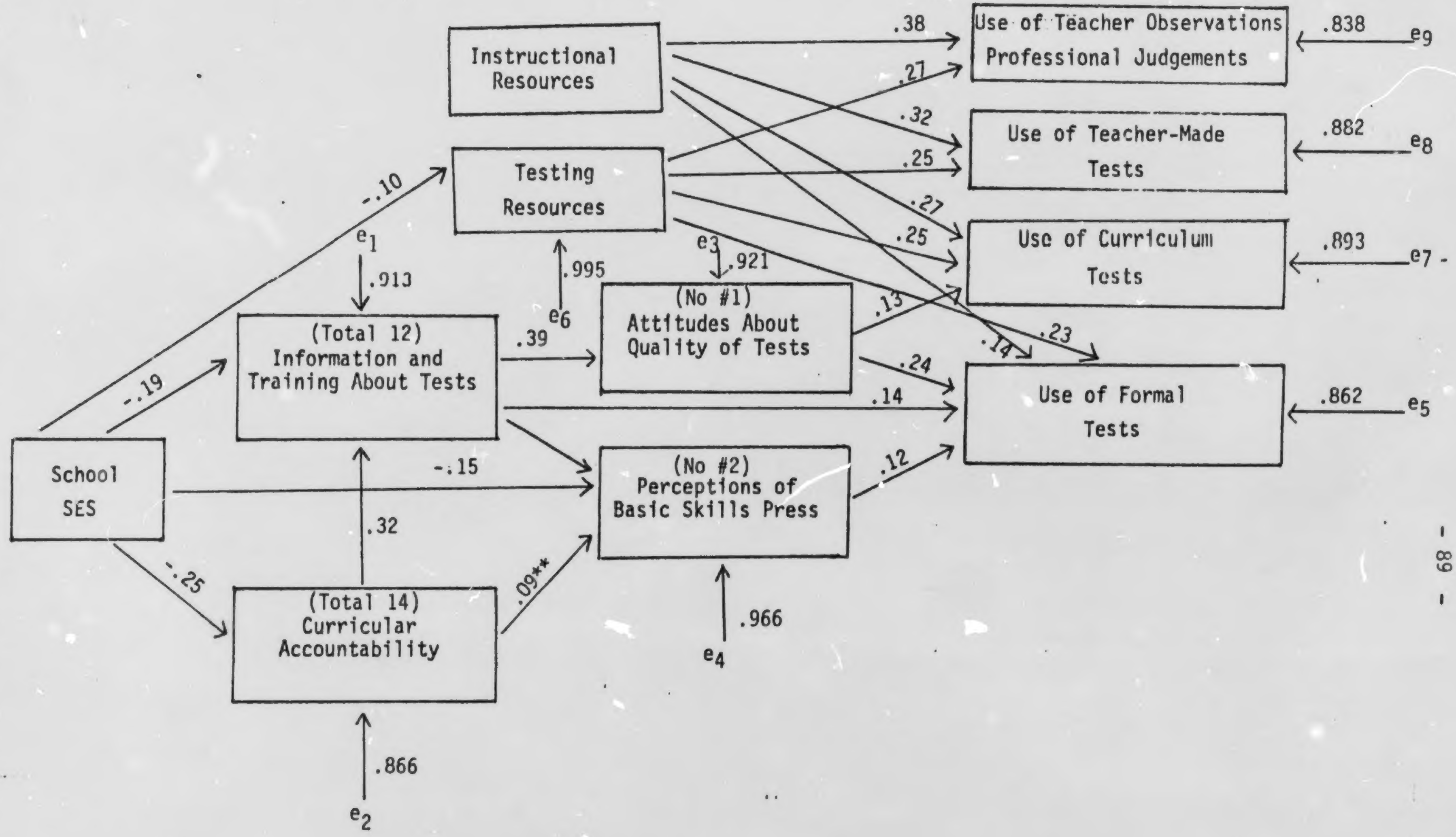

Instructional Resources
Testing Resources
Use of Teacher Observations Professional Judgements
Use of Teacher-Made Tests
Use of Curriculum Tests
Use of Formal Tests
(Total 12) Information and Training About Tests
(No #1) Attitudes About Quality of Tests
(No #2) Perceptions of Basic Skills Press
School SES
(Total 14) Curricular Accountability
.38
.27
.32
.25
.27
.25
.13
.23
.14
.24
.14
.12
.838
e9
.882
e8
.893
e7
.862
e5
-.10
e1
.913
.995
e6
e3
.921
.39
-.19
-.15
.32
-.25
.09**
.966
e4
.866
e2

FIGURE 2

CONCEPTUAL MODEL FOR ELEMENTARY SCHOOL TEACHERS' TEST USE IN MATHEMATICS*

*Reported values correspond to standardized path coefficients that were statistically significant (p .05).

**Reported coefficient statistically significant (p .06).

SECONDARY READING

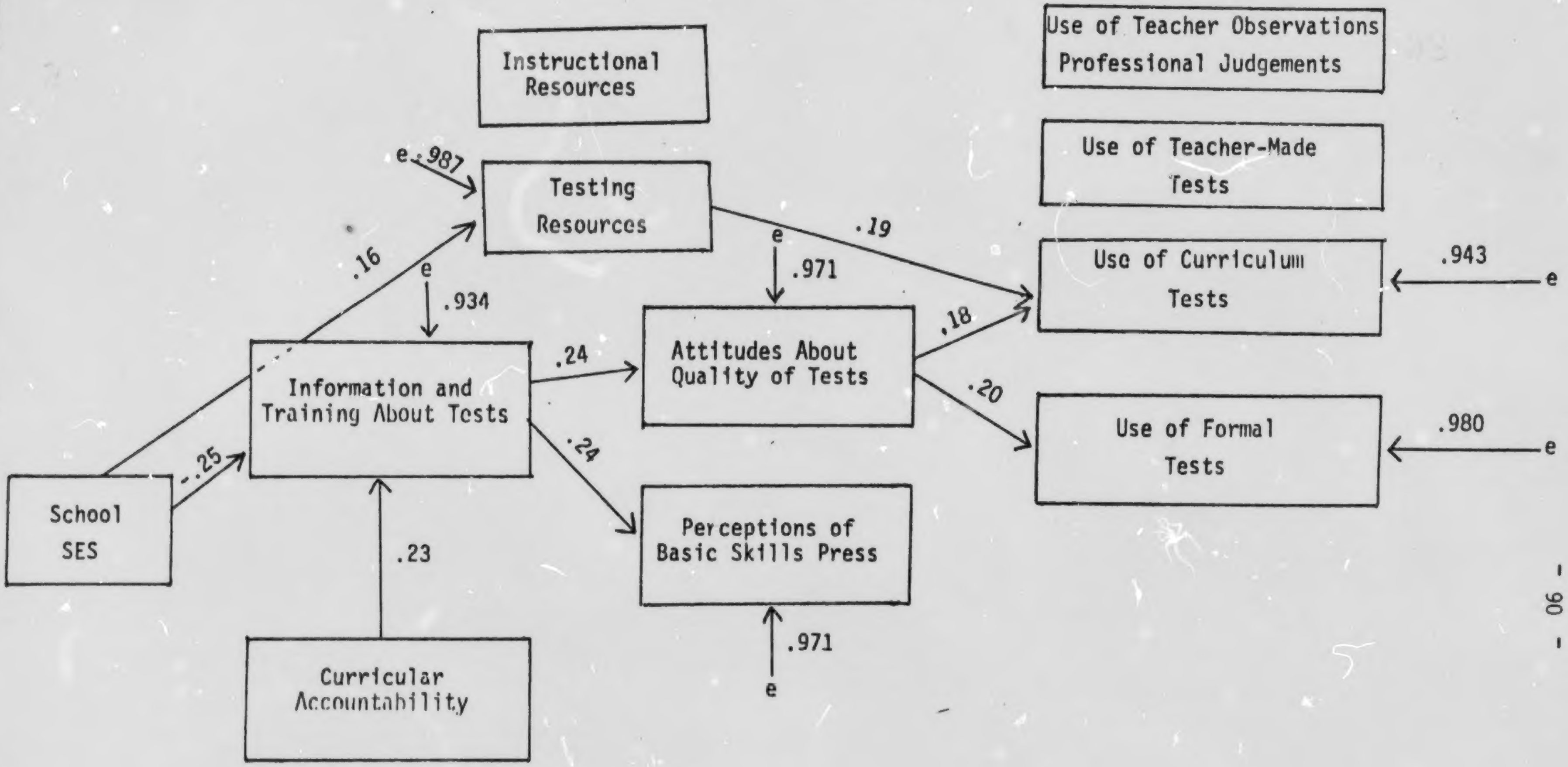

Instructional Resources
Use of Teacher Observations Professional Judgements
e .987
Testing Resources
Use of Teacher-Made Tests
.19
e
.971
.16
e
.934
Use of Curriculum Tests
.943
e
.24
Attitudes About Quality of Tests
.18
Information and Training About Tests
.20
Use of Formal Tests
.980
e
-.25
.24
School SES
Perceptions of Basic Skills Press
.23
.971
e
Curricular Accountability

FIGURE 3

CONCEPTUAL MODEL FOR SECONDARY SCHOOL ENGLISH TEACHERS' TEST USE*

*Reported values correspond to standardized path coefficients that were statistically significant ($p < .05$)

SECONDARY MATHEMATICS

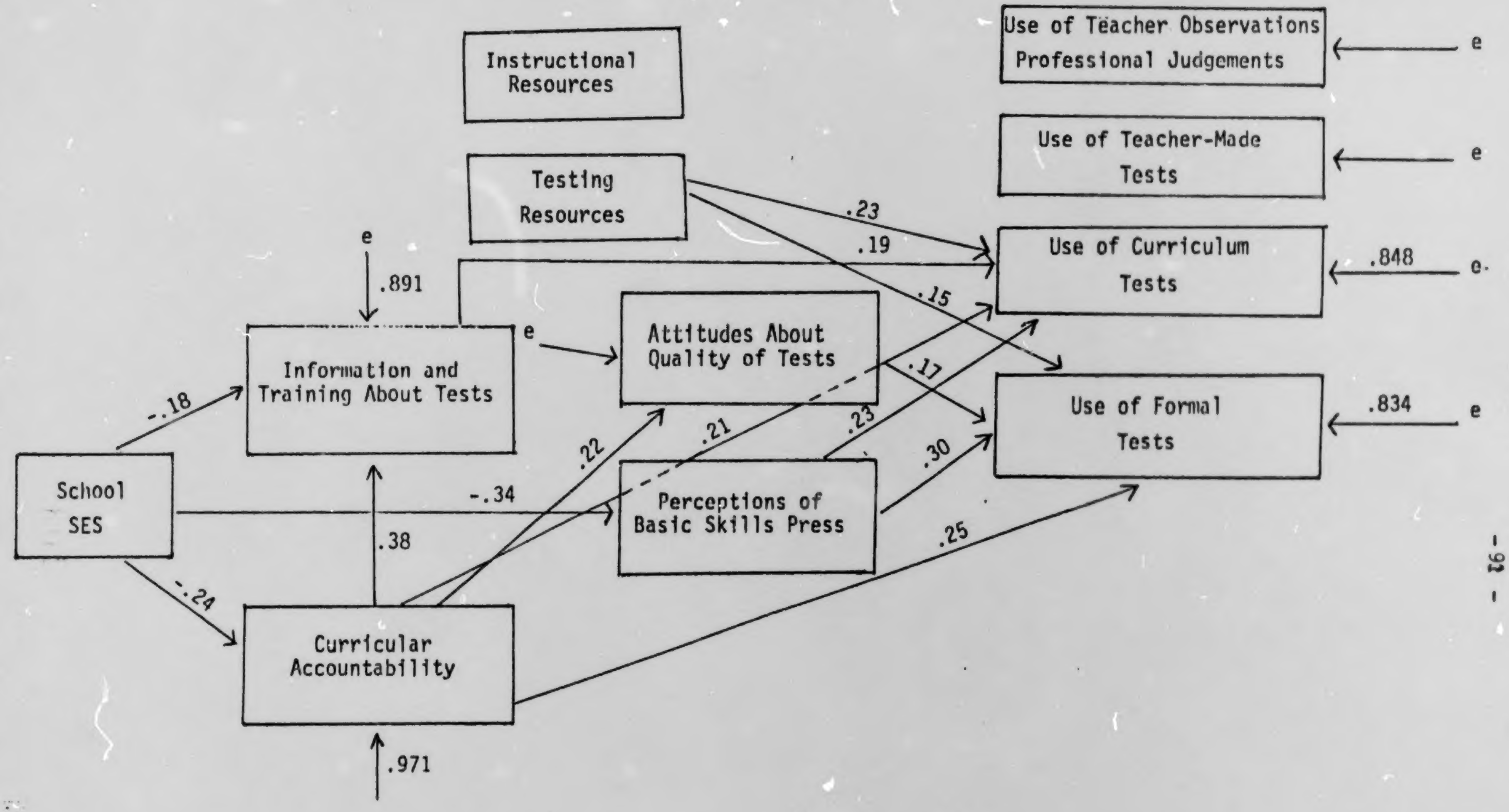

Instructional Resources
Testing Resources
Use of Teacher Observations Professional Judgements
Use of Teacher-Made Tests
Use of Curriculum Tests
Use of Formal Tests
Information and Training About Tests
Attitudes About Quality of Tests
Perceptions of Basic Skills Press
School SES
Curricular Accountability
e
e
e
e
e
e
.23
.19
.15
.17
.23
.21
.22
.30
.25
-.34
.38
-.18
-.24
.891
.971
.848
.834

FIGURE 4

CONCEPTUAL MODEL FOR SECONDARY SCHOOL MATHEMATICS TEACHERS' TEST USE*

*Reported values correspond to standardized path coefficients that were statistically significant ($p < .05$)

References

Airasian, P.W. The effects of standardized testing and test information on teachers' perceptions and practices. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, California, 1979.

Baker, E.L. Achievement testing in urban schools: New numbers. Paper presented at CEMREL Conference on Urban Education, 1978. (Also CEMREL Monograph on Urban Education, St. Louis, Missouri: CEMREL, 1980.

Bank, A., & Williams, R. Evaluation in school districts: Organizational perspectives. CSE Monograph Number 10. Los Angeles: Center for the Study of Evaluation, 1981.

Berman, P., & McLaughlin, M.W. Federal Programs Supporting Educational Change, Vol. VII: Factors Affecting Implementation and Continuation. Report R-1589/7-HEW. Santa Monica: Rand Corporation, 1977.

Boyd, J., McKenna, B.H., Stake, R.E., & Yashinski, J. A study of testing practices in the Royal Oak (Michigan) public schools. Royal Oak City School District, 1975. (ERIC Document Reproduction Service No. ED 117161.)

Center for the Study of Evaluation. CSE Criterion-Referenced Test Handbook. Los Angeles, CA: Center for the Study of Evaluation, 1979.

Choppin, B., & Dorr-Bremme, D. Test use project. Los Angeles: CSE, University of California, Los Angeles. A deliverable submitted to the National Institute of Education, Washington, D.C., 1978.

Dorr-Bremme, D., & Herman, J. Uses of testing in the schools: A national profile. Los Angeles: CSE, University of California, Los Angeles, 1982.

Edmonds, R. Effective schools for the urban poor. Educational Leadership, 1979. (Reprinted by permission of CEMREL.)

Goslin, D.A. The use of standardized ability tests in American secondary schools and their impact on students, teachers, and administrators. New York: Russell Sage Foundation, 1965.

Goslin, D.A., Epstein, R., & Hallock, B.A. The use of standardized tests in elementary schools. Second Technical Report. New York: Russell Sage Foundation, 1965.

Huron Institute. Summary of the Spring Conference of the National Consortium on Testing. Cambridge, Mass.: Huron Institute, 1978.

Resnick, L. & Resnick, D. The social functions of educational testing. A proposal submitted to the Carnegie Corporation of New York, 1978.

Salmon-Cox, L. Teachers and tests: What's really happening? Paper presented at the annual meeting of the American Educational Research Association, Boston, Massachusetts. 1980. (Also Phi Delta Kappan, 1981, 631-634.

Stetz, F., & Beck, M. Teachers' opinions of standardized test use and usefulness. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA, 1979.

Tyler, R. What's wrong with standardized testing. Today's Education, 1977, 66(2), 35-38.

Yeh, J. Test use in schools. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles. Work Unit 4, Studies in Measurement and Methodology, June, 1978.

Testing in the Schools:
Implications of a National Survey of Teachers and Principals

Robert L. Linn
College of Education
University of Illinois at Urbana-Champaign

According to teacher reports obtained from the questionnaire survey conducted by the Center for the Study of Evaluation, (Dorr-Bremme & Herman, 1983), the typical student in grades 4 to 6 or in grade 10 spends a substantial amount of time taking various kinds of achievement tests each year. On the average, it is estimated that students in grades 4 to 6 spend approximately 22 1/2 hours per year taking reading or mathematics tests. Roughly half this time is spent taking tests required by the state or local school district; the other half goes to non-required tests which are selected or constructed by individual teachers.

The corresponding estimate of time that grade 10 students spend taking English or mathematics tests is almost 51 hours per year. The increase, compared to elementary schools, however, is due almost entirely to increases in the amount of time spent on tests selected or constructed by teachers, which accounts for about 38 of the 51 hours, or about 75% of the time.

The observation that quite a few hours are devoted to testing is not particularly suprising. Indeed, the introductory paragraph of the CSE questionnaire seems to presuppose a lot of testing, stating, for example, that "testing firms and curriculum publishers are flooding the system with new test materials".

Of course, knowing that elementary students spend roughly 5% of their reading or mathematics class time taking tests of one form or another, or that the corresponding figure for grade 10 students is closer to 20%, raises more questions than it answers. These results, taken in isolation, do not answer questions of real interest, such as: Should more or less time be devoted to testing? What uses are made of the results? Are the results used appropriately? What are the nature and quality of the tests? Is the balance between classroom and externally required testing about right? What are the positive and negative effects of all the testing? How can testing be used more effectively? Partial answers to some of these and other questions can be gleaned from the CSE survey results. Results of other research studies can also be brought to bear on such questions. Unfortunately, however, we still must rely on rather weak evidence and speculation in trying to answer some of the more important questions.

Moratorium

It is hardly necessary to review here the various criticisms of standardized testing. Some of the common criticisms were mentioned by Dorr-Bremme and Herman, and I'm sure that those and other criticisms are quite familiar to this audience. It is worth noting that those common criticisms are generally directed only at standardized tests, rather than classroom tests which, as was just noted, consume about half of the total testing time at grades 4 to 6 and three-quarters of the time at grade 10. It is clear that some of the more vocal critics not only believe that too much time is devoted to standardized testing, but that any time would be too much.

In this regard, the report of the 1978 National Conference on Achievement Testing and Basic Skills provided the following summary of the position of the National Education Association:

> Since 1971, the NEA has sought a moratorium on standardized testing because of beliefs that the tests do not do what they purport to do, that they tend to be culturally biased, that they automatically label half the students as losers. Standardized tests seldom correspond significantly to local learning objectives, and they can't be used to measure growth over a short period of time... (National Institute of Education, 1979, p. 13).

Judging from the amount of time spent on required testing, it is clear that the extreme action of a moratorium has not hit a very reponsive chord. The call for a moratorium is not only an extreme position, but it seems to conflict with the opinions expressed by teachers in the CSE survey as well as those obtained in several other surveys. In a national survey of approximately 3,300 teachers conducted in 1978-79, for example, Stetz and Beck (1981) found that although about 20% of the 3,140 teachers who responded to the questions said that the amount of standardized testing in the teachers' school systems was "too great", 73% said it was "about right", and an additional 7% said that it was "too little". Goslin (1967) reported similar results for a survey of teachers conducted fifteen years earlier, in 1963-64. Only 15% of the teachers who expressed an opinion in Goslin's survey said they believed that too many standrdized tests are given. Roughly an equal number said that too few were given, and the remaining 68% said the number was about right.

A moratorium, or even a significant reduction in the amount of standardized testing, would seem to be contrary to the stated opinions of a substantial majority of teachers. As will be shown below, the reasons stated for a moratorium also seem to conflict with the opinions expressed by teachers in the CSE and other surveys.

Uses

The CSE survey did not ask teachers if they thought too many or too few tests were given, but from the results of previous surveys it might reasonably be assumed that the modal response would have indicated that the number was "about right". A more important question, however, is what use is made of the test results. The CSE survey asked teachers how important various sources of information were for four purposes: (1) planning teaching at the beginning of the year, (2) initial grouping or placement of students, (3) changing a student from one group or curriculum to another, or providing remedial or accelerated instruction, and (4) deciding on report card grades. Teachers responded on a four point scale of crucially important, important, slightly important, or unimportant. Not suprisingly, standardized tests were judged to be relatively unimportant for the purpose of deciding on report card grades. Actually, the means were a bit higher than I would have expected on this question, falling about halfway between unimportant and slightly important.

The means on the other three uses of standardized tests all fell between slightly important and important. Although these means are lower than the correponding means for teacher-made tests or teacher opinions and observations, I consider this response to standardized tests to be relatively positive. It is certainly more positive than seems to be implied by the previously mentioned NEA position. A number of other studies have found that teachers value their own judgment more highly than the information provided by standardized tests (e.g., Hastings, Runkel, Damrin, Kane & Larsen, 1960; Hotvedt, 1978; Scheyer, 1977; Stake & Easley, 1978). This seems to be a reasonable state of affairs.

As Kellaghan, Madaus and Airasian (1982, p. 259) have pointed out, standardized "test information in most cases serves to confirm the evaluations of pupil ability and achievement that teachers have already formed. Thus, it will be the exception rather than the rule for a teacher to be confronted by information from tests that might lead him or her to believe that some modification of his or her perceptions or practice should be considered." The availability of an independent source of information that identifies such exceptions is one of the important functions that are served by standardized tests.

The range of uses for which teachers were specifically asked to judge the importance of standardized tests in the CSE survey is rather narrow. Neither of the two most frequently reported uses identified in the Stetz and Beck (1981) study are included in the list. Seventy-four percent of the teachers surveyed by Stetz and Beck reported that they used standardized achievement test results for "diagnosing strengths and weaknesses". The second most common use, which was claimed by 66% of the teachers, was "measuring growth". These figures may be compared to 52% for "instructional planning", which is the question in the Stetz and Beck survey that most closely paralleled the CSE importance questions. It would be of interest to know how the CSE respondents would have rated the importance of the other, more common, uses reported by Stetz and Beck.

Minimum Competency Testing

One of the areas that is given more attention in the CSE survey than in earlier studies, such as Goslin's or Stetz and Beck's, is that of minimum competency testing. A majority of the combined sample of elementary and secondary teachers indicated that "tests of minimum competency have affected or would affect the amount of time they could devote to teaching subjects or skills not covered by the tests." Despite this fact, an overwhelming majority, ranging from 81% to 90% in the three groups of teachers surveyed, agreed with the statement that "tests of minimum competency/proficiency/functional literacy should be required of all students for promotion at certain grade levels or for high school graduation". It would have been nice if the promotion and graduation uses had been separated, and if a third category of mandatory assignment to a remedial program had been added.

It is unclear how many teachers favored one of the uses (e.g., promotion) but not the other (e.g., graduation). Judging from the findings of Stetz and Beck, who found that 59% of the teachers favored "the use of competency test results to determine high school graduation", the CSE percentages would probably have been somewhat lower if the uses had been separated. Nonetheless, minimum competency test requirements of one kind or another seem to enojy rather widespread support among teachers.

The level of apparent support is somewhat puzzling when juxtaposed with the same teachers' opinions of the fairness of minimum competency tests. Between 35% and 58% of the three groups of teachers agreed with the statement that "tests of minimum competency are frequently unfair to particular students". If the use and fairness questions are considered together, it must be inferred that at least a third of the teachers simultaneously believe that minimum competency tests are frequently unfair to some students but that nonetheless they should be required of all students for promotion at some grades or for high school graduation. Maybe teachers have faith that their district will avoid one of the tests that are judged unfair. Or possibly they believe that the benefits for most students outweigh the perceived unfairness for a few students. The fact that a clear majority of teachers (between 73% and 93%) say that testing motivates students to study harder may help explain the apparent inconsistency. But I still find these opinions rather difficult to reconcile.

It is worth noting that the teachers from schools with minimum competency testing requirements had somewhat less favorable attitudes toward this use of tests than did their counterparts from schools that did not have a minimum competency testing program. Given the external pressure on teachers for accountability, it may be that teachers believe that it is prudent to accept such a requirement in principle. But experience with the limitations of an actual program may dampen their enthusiasm.

The apparent strength of the endorsement of minimum competency test requirements also seems a bit suprising when coupled with the previously mentioned finding that most teachers think that such requirements would alter the amount of time that they would devote to content or skills not covered by the test. The latter opinion certainly seems reasonable. There is considerable evidence that examinations that have important consequences do influence the curriculum (see, for example, Cronbach, 1963; Linn, 1983a,b; Madaus & Greaney, 1982; Madaus & McDonagh, 1979; Tinkleman, 1966). But one of the more common criticisms of minimum competency testing is that it will narrow the curriculum, and I would have expected that teachers would resent the shaping of the curriculum by such an external force.

Of course, while some people see the prospect of a test-driven curriculum as a danger, othes see it as a desirable end and would argue that a "test provides the means of making agreed-upon objectives clear and precise. An important goal of instruction should be the achievement of those objectives as demonstrated by performance on the test" (Linn, 1983a, p. 125). Nonetheless, I find it a bit surprising that teachers are apparently so sanguine about having an external test play such an important role in determining what they teach.

The "Debra P." case has made it clear that students must be provided with instruction in the content and skills covered by a test that is required for high school graduation. Instructional validity was a central issue in that case and can be expected to be a key consideration in other judicial decisions regarding minimum competency tests. The 1981 decision of the Fifth Circuit Court of Appeals concluded that "A state may condition the receipt of a public school diploma on the passing of a test so long as it is a fair test of that which was taught" (644 f.2d at 406). Because the Court of Appeals did not find sufficient proof in the record before it in 1981 that "the test covered material actually studied in the classrooms of the state", the case was remanded for further findings. Subsequent to that decision, Florida commissioned IOX Assessment Associates to conduct a massive study of the instructional validity of the test.

That study consisted of a survey of teachers, a survey of school districts, a survey of students, and a series of site visits. Over 25,000 elementary and secondary comunications teachers and a similar number of mathematics teachers responded to the teacher survey. For each of the 24 skills tested on the State Student Assessment Test, Part II (SSAT-II), teachers were asked to answer the following question:

> "During the previous instructional year, did you provide instruction which specifically prepared your students for this SSAT-II skill?"

Those who answered yes to this question were asked to respond to a second question:

> "Did you provide your students with sufficient instruction so that they should be able to demonstrate mastery of this skill on the SSAT-II?"

Needless to say, the teacher survey alone, not to mention the other three components of the study, produced a voluminous amount of data.

Although I have reservations about the results for demonstrating the instructional validity of the test that I expressed in testimony before the District Court, I won't go into that issue here. My only reason for describing the study is to underscore the importance of the match between what is taught and what is tested when a minimum competency test is used to determine the award of high school diplomas. I should note in passing, however, that the District Court was convinced by the results of the IOX study and concluded that the

State had succeeded in "proving by a preponderance of the evidence that the SSAT-II is instructionally valid and therefore constitutional". Whether that decision will stand following appeal remains to be seen, but the state of Florida was allowed to deny diplomas to students in the class of 1983 who had not passed the test.

The CSE survey provides only meager and somewhat ambiguous information on the question of instructional validity. Teachers were asked the degree to which they agreed or disagreed with the following statement: "The content (or skills) of most required tests is very similar to the content or skills that I teach." Note that minimum competency tests are not singled out, and that such tests comprise only a small fraction of the required tests. Nenetheless, the responses of the teachers are of interest in light of the importance of this issue and the previously quoted position of the NEA that "standardized tests seldom correspond significantly to local learning objectives".

The responses of the teachers to the CSE survey are contrary to the NEA claim. Slightly over three-fourths of the teachers agreed that the content of the tests was very similar to that which they teach. It is, of course, important that between one-fifth and one-quarter of the teachers disagreed with the statement. When it is considered that the question covers a wide range of tests, however, the results, along with those in Florida, would seem to provide encouragement to those who hope to demonstrate that teachers consider a carefully selected test of minimum competency to have instructional validity.

Although it again mixes minimum competency tests with other standardized tests, one other item on the teacher questionnaire that deals with minimum competency testing is worthy of mention. This is the question of whether teachers should be held accountable for students' scores on these tests. Not surprisingly, a substantial majority of between 61% and 71% of the teachers said, "No." Principals apparently concur. Or at least they generally rated the importance of minimum competency tests for purposes of evaluating their teachers as either "unimportant" or "slightly important".

Secondary school principals and, to a lesser extent, elementary principals, gave relatively high importance ratings to the information provided by minimum competency tests for several uses other than teacher evaluation. Interestingly, for both groups of principals, five uses of the information were rated to have greater importance than deciding whether to retain or promote students, including deciding whether a student should graduate or receive a certificate. The latter use had an average rating about halfway between slightly important and important. The five uses that received higher ratings of importance, in order of their average ratings from secondary school principals, were:

1. deciding what areas of the curriculum need added or reduced emphasis (rated 3.27)
2. reporting to district personnel about the academic progress or problems of the principal's school (rated 3.12)
3. communicating to parents about their child's progress or problem (rated 3.03)

4. assigning students to classes (rated 2.98)
5. informing the public (e.g., through the newspaper, at meetings, etc.) about the academic progress or problems of the principal's school (rated 2.92).

For secondary school principals, the order of the ratings for the five just-mentioned uses is identical for standardized tests and for minimum competency tests, but the latter type of tests received a slightly higher average importance rating in each case. Of greater interest is the fact that the results of minimum competency tests are rated to have somewhat greater importance than are teacher opinions and recommendations, or teacher-made and curriculum tests, for three of the above uses. The sources of information are rated of equal importance for a fourth purpose, assigning students to classes. Only for the purpose of communicating to parents are teacher opinions and test results rated as more important than minimum competency tests, and here the latter source of information has a mean rating of important (3.03), with a standard deviation ranging from slightly important to crucial.

The secondary school principals seem to attach a good deal of importance to minimum competency and other kinds of standardized tests for these five particular purposes. The ratings of elementary principals are lower, but they still indicate that the results of these two types of tests are fairly important for these five purposes.

Standardized Tests

Three of the uses of standardized tests that were rated for their importance by secondary school principals in the CSE survey have close parallels in Goslin's (1967) questionnaire that was given to secondary school administrators in the early 1960's. Goslin also used a four point scale ranging from "no importance" to "very important", which is similar, albeit not identical, to the CSE scale. A comparison between the mean ratings on the three similar items in the two questionnaires is shown in Table 1.

Given the difference in the labels attached to the scale points in the two studies, and the slight differences in the wording of the questions, exact comparisons of the two sets of results are not possible. It would appear, however, that the importance attached to standardized test results for class assignments has increased, that for curriculum evaluation has remained about the same, and that there has been some decline in the importance for teacher evaluation.

None of the three reasons for using standardized tests that secondary school administrators considered to be of greatest importance in the Goslin study were considered in the CSE study. Those uses were "to help pupils gain a better understanding of their strengths and weakness" (mean rating of 3.68), "to help in educational and vocational counseling of pupils" (mean rating of 3.66), and "to help in guiding pupils into appropriate curricula" (mean rating of 3.37).

For elementary school principals, the comparison of the CSE results to those obtained by Goslin is less direct. In his study, Goslin asked principals to list up to four main uses for several types of tests. For both reading and arithmetic achievement tests, elementary school principals listed two uses on the average. Slightly over three-fourths of the principals said that "diagnosing learning difficulties" was one of the main uses of both standardized reading and arithmetic tests. For reading, the second and third most commonly mentioned uses were homogeneous grouping, listed by 42% of the principals, and curriculum evaluation, listed by 32%. The same uses were also the second and third most commonly mentioned uses for arithmetic tests, but with the order reversed. Thus, two of the three most common uses identified by elementary principals in the Goslin survey were both included in the CSE survey.

Technological Aides and Staff Development

The remainder of my comments on the CSE survey wil be focused on two topics that have not previously been touched upon. These are staff development related to testing and the availability and use of technological resources. Table 10 of the Dorr-Bremme and Herman report lists the percentage of teachers reporting participation in various staff development activities. The most frequent participation is in areas that might be characterized as being more administrative in nature, e.g., analysis and explanation of state, district, or school results, or how to adminsister required tests. At the other end of the continuum are activities that appear to be more instructionally related, e.g., how to construct or select good tests and the use of test results to improve instruction.

This distribution seems to bear almost an inverse relationship to the needs and priorities of educators. Both of the quotations that Dorr-Bremme and Herman gave from their interviews of teachers emphasized instructional uses of test results and the desire for information that would help in diagnosing difficulties and prescribing instruction. As I've already indicated, both the Goslin and the Stetz and Beck surveys yielded results that underscore the importance to educators of test results that will help in diagnosing a student's strengths and weaknesses. The close linking of testing and instruction is certainly an understandable goal, but not one that is easily accomplished. I believe that there are good reasons for thinking that greater emphasis, both in the area of staff development and in the development and use of techological resources, is needed in order to realize the goal of making better instructional use of tests.

The instructional use of tests requires more than global scores. A low score on a standardized arithmetic test, for example, signals a problem, but by itself does not identify the nature of the problem or indicate what should be done about it. More fine-grained information about clusters of items that measure a common skill is needed. Better yet, the nature of the error that is consistently made on particular types of problems needs to be identified so that it can be corrected. Davis (1979, p. 6) has noted that "one of the most common student requests is, 'Tell me what I am doing wrong.'"

Perceptive teachers can often meet this student request, but it requires careful attention, not only to whether the stduent gets the right or wrong answer. They must determine what kind of error was made and whether it represents a systematic misconception or erroneous algorithm in order to be fully responsive to the student's request for help. In the last few years, there has been an increasing number of studies that have demonstrated that student errors are generally "not random or careless, but...driven by some underlying misconception or by incomplete knowledge" (Glaser, 1981, p.926).

Brown and Burton (1978) have referred to the misconceptions that often lead to the systematic errors that are made by students as "bugs". They and several other researchers (e.g., Bartholomae, 1980; Davis, 1979; Davis, Jockusch & McKnight, 1978; Siegler, 1978; Tatsuoka, 1982) have demonstrated that such bugs are quite common. For example, Tatsuoka has identified specific types of errors that are made systematically by some students in arithmetic operations with signed numbers. Once a particular type of error has been diagnosed for a student, his/her answers on other problems can be predicted with very high accuracy. More importantly, pinpointing the precise nature of the error is half the battle in getting it corrected.

Error analysis has great potential for improving instruction. But it requires considerable skill and effort in the construction of test items that can distinguish among various misconceptions. It also requires a different level of analysis of the responses than just computing number of right scores. Staff development and ready access to resources that can ease the burden of item development and analysis are needed to take advantage of the potential.

The CSE survey results suggest that when teachers have access to resources, such as item banks and quick, computerized scoring and analysis of tests, they make considerable use of them. Unfortunately, less than half of the teachers report that these relatively straightforward resources are available. However, these are functions that can be readily served by a microcomputer, and it seems reasonable to expect that access to micros will soon become commonplace. Indeed, in many schools it already has.

Hsu and Nitko (1983) have recently reviewed some of the current and potential uses of micros for various educational testing functions. Though not intending to be comprehensive, they identified

31 software packages, ranging in price from $15 to $300, that are currently available on various micros such as the Apple II. The functions served by these packages range from item banks, item analysis, and test scoring on the one hand, to on-line testing, diagnostic testing, and adaptive testing on the other.

The technical capability to support and improve classroom testing exists. Effective utilization will require a considerable developmental effort, however. User-friendly systems and teacher guides, such as the one being developed under support from NIE by Nitko and Hsu, are essential. But the potential payoff for the effort could justify the cost many times over.

Conclusion

The CSE survey has given us a glimpse at the use of tests in elementary and secondary schools. Many of the results are similar to those from earlier surveys. Teachers and principals say they use the results of tests and attach more importance to them than is generally claimed by test critics. As would be expected, teachers' primary interest is in results that have direct instructional value by identifying the strengths and weaknesses of individual students, and they generally rely more on their own tests and observations than on standardized tests for these purposes. The availability of resources that support the development and use of tests for this primary instructional purpose is limited. However, microcomputer technology has the potential to radically alter the situation. If properly developed, the instructional value of testing could be greatly enhanced by making better use of this technology.

Table 1

A comparison of the mean ratings of the importance of three similar uses of standardized tests in the Goslin (1967) and Dorr-Bremme and Herman (1983) surveys of secondary school administrators.

Use	Goslin[1]	Dorr-Bremme & Herman[2]
Curriculum Evaluation	3.05	2.91
Student Class Assignments	2.33	2.77
Teacher Evaluation	2.28	1.63

1. Scale: 1 = of no importance, 2 = of very little importance, 3 = fairly important, 4 = very important

2. Scale: 1 =unimportant, 2 = slightly important, 3 = important, 4 = of crucial importance

References

Bartholomae, D. (1980). The study of error. College Composition and Communication, 31, 253-269.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. Cognitive Science, 2, 155-192.

Cronbach, L.J. (1963). Course improvement through evaluation. Teachers College Record, 64, 672-683.

Davis, R.B. (1979). Error analysis in high school mathematics, conceived as information processing pathology. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Davis, R.B., Jockusch, E., & McKnight, C. (1978). Cognitive processes in learning algebra. The Journal of Children's Mathematical Behavior, 2, 1-320.

Dorr-Bremme, D., & Herman, J.L. (1983, July). Testing in the schools: A national profile. Paper presented at the Center for the Study of Evaluation conference "Paths to Excellence: Testing and Technology", Los Angeles, UCLA.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Goslin, D.A. (1967). Teachers and testing. New York: Russell Sage Foundation.

Hastings, J.T., Runkel, P.J., Damrin, D.E., Kane, R.B. & Larsen, G.L. (1960). The use of test results. Urbana, Illinois: University of Illinois, Bureau of Educational Research, 1960.

Hotvedt, M. (1978). Teacher uses of testing. Paper presented at the J. Thomas Hastings Symposium on Measurement and Evaluation; Urbana, Illinois, University of Illinois.

Hsu, T., & Nitko, A.J. (1983). Microcomputer testing software teachers can use. Paper presented at the 1982 ECS Large-Scale Assessment Conference, Boulder, Colorado.

Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1982). The effects of standardized testing. Hingham, MA: Kluwer-Nijhoff.

Linn, R.L. (1983a). Curriculum validity: Convincing the courts it was taught without precluding the possibility of measuring it. In G.F. Madaus (Ed.), The courts, validity and minimum competency testing. Hingham, MA: Kluwer-Nijhoff.

Linn, R.L. (1983b). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20, 179-189.

Madaus, G.F. & Greaney, V. ((1982). Competency testing: A case study of the Irish Primary Certificate Examination. Occasional Paper Series, Cambridge, MA: National Consortium on Testing.

Madaus, G.F. & McDonough, J. (1979). Minimum competency testing: Unexamined assumptions and unexplored negative outcomes. In R. Lennon (Ed.), Impactive changes in measurement: New directions for testing and measurement, 3(3), 1-14.

National Institute of Education. (1979). The National Conference on Achievement Testing and Basic Skills. Washington DC: National Institute of Education.

Scheyer, P. (1977). Test results revisited. The Cornbelt Education Review: A Graduate Student Journal. Urbana, IL: College of Education, University of Illinois.

Siegler, R.S. (1978). The origins of scientific reasoning. In R.S. Siegler (Ed.), Children's thinking: What develops? Hillsdale, NJ: Erlbaum.

Stake, R.E., & Easley, J.A., Jr. (1978). Case studies in science education, booklet XIII, findings II. Urbana, IL: Center for Instructional Research and Curriculum Evaluation and Committee on Culture and Cognition, University of Illinois.

Stetz, F.P., & Beck, M.D. (1981). Attitudes toward standardized tests: Students, teachers and measurement specialists. Measurement in Education, 12(1), 1-10.

Tatsuoka, K.K. (1982). Effect of different instructional methods on error types and their consistency and change at different points in learning. (Research Report 82-5-ONR) Urbana, IL: University of Illinois, Computer-based Education Research Laboratory.

Tinkkleman, S.N. (1966). Regents examinations in New York State after 100 years. Proceedings of the Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service.

Conceptions of Testing In the Public School

Robert Calfee

School of Education

Stanford University

In preparing this review of the survey of test use in the schools by the Center for the Study of Evaluation, I was mindful of several recent newspaper articles in which tests figured prominently:

As reported in A Nation at Risk, the nation's schools have declined in quality to crisis proportions; tests are one of the primary sources of data in support of this claim.

Even though they serve students from poor families and may lack adequate financing, some schools appear to excel. At Pioneer High School in southern California, for instance, none of the students fail to graduate because they cannot pass the district's minimum competency test, even though the school is in a poor and predominantly Hispanic neighborhood.

President Reagan has recently suggested an initiative to raise SAT scores nationwide by 50 points--both verbal and quantitative, so I understand.

According to SB 813, the educational reform legislation just passed in California, if individuals have a bachelor's degree and can pass two state tests, then they will be certified as high school teachers after a two-year apprenticeship in any of the state's districts.

Each of these examples demonstrates the significant role of achievement tests in the arena of educational politics, a role that has reached substantial proportions over the past few decades. Why should politicians be fascinated by tests? Why are politicians fascinated with anything? The answer is often power. And tests do constitute a source of power, a lever that can change educational practice for better or worse, which has considerable appeal to legislators, bureaucrats, judges, and various special interest groups, as well as school administrators. In addition to being a source of power, tests are relatively cheap and can be centrally controlled--an attractive combination.

"Tests" as defined by the preceding context take on a well-known configuration: a group-administered, multiple-choice, paper-and-pencil task, usually designed to assess "basic skills" (reading and arithmetic), usually designed and developed by an agency that is external to the classroom and the school.

This CSE conference interrupts a vacation by my wife and me at Carmel, where we are attending the Bach festival and the Master's Festival at Hidden Valley Music Ranch. On Monday morning, we were privileged to attend a master's class where several of the most promising young flutists in the world performed for Julius Baker and Jean-Pierre Rampal. I say "a class", and yet there was little obvious "teaching". Instead, each candidate played a selection while the two masters listened. Occasionally, the masters would interrupt with a comment, critique, or suggestion--in fact, the session was a marvelously engaging, informative (and stressful) test!

What a different conception of testing, compared to the student sitting alone at a desk filling in the spaces on a multiple-choice test! To say that the master's assessment was "performance-based" misses the point; the setting, the standards, the scoring--on each of these dimensions and others that might be explored--the master's session was virtually nonoverlapping with the conception in the public mind. To be sure, whether one conception is better than the other depends on one's purposes and values.

In any event, I come to this discussion of the CSE test use survey with a relaxed mind and broadened perspective. In the time available, I will address the following three questions:

What can be proposed as a workable conception of achievement testing for public education in the United States?

What operational definitions of achievement testing are of greatst importance in the schools today, viewed against the theoretical perspective provided in answer to the first question of what the "realities" are?

What purposes (and concomitant audiences) are served by the various operational definitions?

Complete and thoughtful answers to these three questions would clearly take me beyond my mandate and my resources. In answer to the first question, I will sketch a framework that I found helpful in organizing my thoughts about the survey. The bulk of the paper will be devoted to the second question; I will present a review and critique of the CSE survey, and then suggest what I think can and cannot be learned from this data set. Finally, I will put forward some opinions in answer to the third question, opinions that take the form of cautions and recommendations for research and practice. In preparing this paper, I have drawn from the material in Testing in the Schools: A National Profile (Dorr-Bremme & Herman, 1983), which was given to all speakers at the symposium, as well as Annual Reports describing the activities of the Test Use and Evaluation Design

Projects (Dorr-Bremme, Choppin, & Burry, 1981; Bank & Williams, 1981).

Conceptions of Achievement Testing

What is the concept of achievement testing that provides the foundation for the test use survey? This issue is not addressed in the Dorr-Bremme and Herman paper, nor is it obvious from the survey instruments for the teachers and principals. In some of the pilot studies that preceeded the national survey, teachers were asked to talk about what they thought should be included under the rubric of testing. Dorr-Bremme et al. (1981, p. 32) present some interesting insights from their discussions with teachers during this pilot work:

> "...respondents referenced [assessment techniques] almost always by their proper names or by vernacular variants of proper names. That is, they rarely talked about norm-referenced tests, criterion-referenced tests, objectives-based tests, curriculum-embedded tests, etc. Instead, they talked about the Ginn placement, the CTBS, the Key Math, 'that state matrix test', and so on..., [or] they gave them functional class names, e.g., diagnostic tests, placement tests, pre-tests, semester finals, 'the competency tests', [and so on]."

By relying primarily on concrete or functionally descriptive titles, practitioners reveal that (a) they are performing practical tasks in the workaday world (likely), or (b) they do not have a separate technical language to describe testing (also probable), or (c) both (most likely, in my opinion). In any event, it appears that the terms of the academic testing profession (NRT, CRT, DRT, etc.) are not catching on in the world of practitioners. Dorr-Bremme et al. (1981) also note that teachers include in their list of testlike things such entries as "homework, worksheets, conferences, book reports, discussions, observations, [inter alia]." (p. 33)

Any conception of achievement testing begins with the notions of collecting evidence for the assessment of what a student has learned in school--what he or she knows, and how well the knowledge can be applied. By achievement, I assume that we are referring to school achievement, so that knowing and doing are both important.

Within this general constraint, I would propose that the folowing dimensions are important facets of the overall concept of achievement testing:

What is tested? The subject matter, the assessment of what has been taught, but also how well what has been learned can be applied in other contexts, the establishment of standards--all of these would be placed under the "what' rubric. Ralph Tyler, speaking at AERA this past spring, described what he thought were the major achievements to be attained by students as a result of their educational experiences in our public schools. I have not yet found time to transcribe his remarks, but let me simply suggest that his answer to "what", viewed as one individual's ideal, provides an interesting framework for consideration of the lists of objectives that are encountered elsewhere.

How to test? I will not elaborate on this dimension, other than to remind you of the contrast between the high school student working through the list of multiple-choice questions that may determine the award of a high school diploma (questions on content that may not have been covered in any of the student's school courses), and the master's class described earlier in the paper--and the numerous variations in "how" that fall between (and beyond) these extremes.

When to test? The spring, and to some extent the fall, are the times when a great deal of emphasis is placed on testing. Test scores in the spring measure the year's learning for the annual report to the Board. Fall testing is for student placement, or for "pretests" if categorical programs are to be evaluated. These times are convenient for some purposes and not for others. Teachers are seldom inclined to use "cold" data.

In California, a major change in "when" has just been legislated with regard to state assessments: first grade testing has been eliminated, and testing at eighth and tenth grades will be added to the previous assessments at third, sixth, and twelfth grades. Competency tests for high school certification are popular throughout the country; the idea is that, before receiving a diploma, it is important to determine that the student is minimally literate in reading and arithmetic. These tests are often administered in the tenth grade or later, with major disruption in the high school program if the student fails. I have suggested elsewhere that these programs are the wrong kind of test at the wrong time and for the wrong purpose; better to ensure minimal literacy before entry to high school. Timing, in any event, is a critical dimension to achievement testing.

Who to test? This question, which also serves primarily as a placeholder, may seem rather strange at first glance. The public image is probably that all students are tested. In fact, not all students are tested in the same way. Students may be absent, and not as a consequence of random events. Learning-disabled students receive different tests designed for different purposes. LES/NES students may or may not be tested. The SAT and the College Board tests are taken by selected groups of students. It may appear that all students are included in the statewide California Assessment Program battery; determining the actual sample that is included from the reports from the twelfth grade testing would be an interesting research project.

Why to test? Other scholars have explored this question (e.g., Cronbach, Anastasi, and so on). The major purposes include selection, assignment, certification, diagnosis, and monitoring,among others.

Finally, for whom to test? Occasionally a student may decide to test himself or herself. Achievement testing more often takes place to meet the needs of someone besides the student: the teacher, the principal, the district, the state, and so on. On occasion, it appears that testing is a routine established at an earlier time for forgotten purposes, kept in place through inertia, without any clear audience.

So there you have it: my representation of the semantic space that defines the overall conception of achievement testing. Within this space one finds many alternate conceptions. By framing the space as I have, it is possible to compare and contrast various alternatives. The framework also helps in mundane matters like designing questionnaires, analyzing data, and interpreting the results of such analyses.

The CSE Survey of Test Use

What does the CSE survey say about the present state of affairs as regards the role of testing in the public schools from the perspectives of principal and teachers? I have organized my thoughts on this matter into four categories:

The guiding questions behind the survey.
The characteristics of the survey instrument.
The sample of respondents.

The results of the presently available data analyses.

The Guiding Questions

Five questions are listed by Dorr-Bremme and Herman; two additional questions were presented to the respondents in the survey information sheet. Here is my effort to bring these questions together into a common framework:

What kinds of achievement testing take place in the nation's schools?

What are the costs (especially in time) of these activities?

What are the perceived benefits to practitioners (the teachers and principals in the study) of the various kinds of testing?

How is the information used?

What administrative practices serve to direct and support various kinds of testing activities?

What are the "perceptions" (opinions, feelings, and so on) of practitioners about various aspects of testing?

What factors are correlated with variation in the responses given to the preceding questions?

I trust that this amalgam is a reasonably accurate reflection of the intentions behind the survey; the questions driving any complex project tend to change over time. Indeed, revision and refinement of questions can be one of the most important outcomes of a research project, outcomes that unfortunately may not be appreciated by funding agencies.

In any event, let me note that this list of questions matches only in part the framework that was sketched in a previous section of the paper. In particular, it appears to be taken for granted in the survey that "everyone knows" what is meant by achievement testing.

Variations in the respondents' underlying conceptions of testing may have influenced their answers, but there is no indication that the identification of individual conceptions of testing was among the primary purposes of the survey.

The Survey Instrument

In reviewing the questionnaires, I found it difficult to discover the themes and concepts that were central to the design of the instrument. I could impose various organizing principles of my own, of course, but the instruments did not "hit me in the face" with categories. In this respect, these instruments resemble the achievement tests that were the focus of the survey. Participants were given a general idea of the topic to be covered, but then, with few exceptions, the questions were presented in a list structure, an organizational structure that is poorly suited to the characteristics of the human mind. To be sure, it may be that a respondent's answers do not depend on whether the format is a list or a set of organized chunks. I know of no research on this question. I do know my own personal reaction when attempting to complete a questionnaire that does not give me a clear picture of where I am being led.

More to the point, the absence of a clear and coherent organizational framework can lead to problems in the construction of an instrument. In the present instance, the teacher and principal instruments fail to mesh at several critical points. It is not a matter of imposing an exact match; one does not want to ask the two groups exactly the same questions. However, it is both possible and of some importance to ensure that the same points are covered whenever possible. A more explict overarching design would have made it easier to compare and contrast responses for the members of a school staff.

Several of the tables in the Dorr-Bremme and Herman (1983) report illustrate the problem.

A couple of minor asides. The survey provided relatively little opportunity for respondents to report their perceptions of the negative impact of tests, which means that the overall tone of the findings may be more positive than would have otherwise been the case. Second, and somewhat related, no space was provided for comments; these can pose problems for analysis and reporting, but they can also provide useful contexts for interpretation of "hard data".

The Sample of Respondents

It appears that the CSE staff did a good job of identifying the sample for the survey. The lattice approach is elegant and efficient, and well suited to the present problem. The five dimensions listed by Dorr-Bremme et al. (1981) as the basis for the sample include four demographic factors (region of the country, metropolitan status, SES and size), and one test-related factor (status with regard to minimum competency testing). The reliance on an efficient design, which I heartily endorse, might have yielded even higher payoff if other factors related to district test policies had been included in the design.

Problems in response rate were described by Dorr-Bremme et al. (1981). The return rate (approximately 60 percent) is troubling only to the degree that the survey purports to represent a "national profile". The fact that no primary teachers were surveyed also limits the generality of the results to some degree. As long as the reader is apprised of these limitations, they do not seem to me to be of major consequence.

The Findings

Before commenting on the results presented in the various reports on the survey, it is important to note that a considerable amount of information has not yet been presented. For instance, the teacher questionnaire provides background data on the respondents, lists of the specific tests used by teachers (difficult to analyze, but given that teachers rely on "names", a rich and important data source), and several subquestions on test use (e.g., are test results returned in a timely fashion) that apparently have yet to be analyzed. In the principal questionnaire, there are data on the school characteristics (not on the background of the principals!),on grouping practices, along with with lists of tests and several subquestions on test use that are unreported. The Center plans to analyze the results for school cohorts (principal and teachers from a given school); these analyses should be of considerable interest. Finally, the data available are mostly means and occasionally standard deviations. It would be helpful for descriptive purposes to know what some of the distributions look like, especially given the categorial nature of most of the responses.

With these caveats in mind, here are highlights of the findings that struck my eye (I assume that the reader of this paper has access to the Dorr-Bremme and Herman report, and the tables therein). The highlights are organized in terms of the six questions listed earlier.

What kinds of tests are used?

These data are available, but results are not reported in the presentation of the findings.

What are the costs?

There is actually some implicit information available about the "kinds" of tests in the results on the costs. It appears that at both the elementary and secondary grades, approximately three hours per year is spent for state-mandated testing of reading and another three hours for mathematics; an equal amount of time is spent in district mandated testing. These data, averages over the entire sample of schools, districts, and states, probably entail testing of the group-administered multiple-choice variety, testing that principals and/or teachers perceive as externally mandated. Given the assumption that an hour per day is spent on reading and on mathematics at all grades, these findings suggest that testing for external purposes takes up about six of the 180 days during the school year. At the elementary level, the report is that another six hours per year are spent in "nonrequired" testing; at the secondary level, the report is about 20 hours of additional testing. These numbers are averages, and one suspects that there is considerable variability between schools and districts.

The validity of these reports deserves scrutiny. For instance, Dorr-Bremme and Herman (1983) mention that there was some confusion about the meaning of "required". It appears that some teachers interpreted this label to refer to tests that were mandated by the teacher for instructional purposes such as grading. In addition, Dorr-Bremme et al. (1981) talk about the "transparency" of everyday activities. Elementary teachers routinely assign worksheets and other testlike activities, and may overlook these in estimating the amount of testing that takes place. High school teachers are more likely to identify testing events in a relatively clear manner; "Put your books away, take out a piece of paper, we're going to have a test."

What are the benefits; how is the information used?

The data show that both principals and teachers agree on one point: externally mandated tests are less useful than teacher-made and curriculum-embedded tests for many purposes. Moreover, tests of any sort are viewed as less informative than nontest data (including teacher judgment) as a basis for decision-making. The basic data are displayed in Tables 4 and 5 in Dorr-Bremme and Herman, 1983. Decisions of central importance to instruction and achievement, including grades, placement, group assignment, and promotion, are perceived as primarily dependent on teacher judgment, secondarily dependent on class-level tests, and least dependent on external tests. In only two areas (data from principals) does it appear that this pattern does not apply: reports to the district and "public information " are based on external test findings more than teacher judgment.

What administrative support and direction are provided for guiding assessment?

Principals (and district administrators) provide substantial assistance to teachers in the area of assessment, far more assistance than teachers report they receive. Teachers say that they are told how to administer tests, and that they are given the results, but otherwise they report that they receive little aid. Administrators and teachers concur on two points. First, achievement testing is not used in any admissable way for teacher evaluation. Second, specific test standards are seldom established for individual schools.

What are practioners' perceptions of testing?

Teachers report that most "required" tests measure what they teach. "Required" may refer to tests that they mandate, rather than externally-mandated tests. Teachers feel that tests are a means of motivating students to do better, and they support competency tests--both elementary and secondary teachers--and overwhelmingly so. However, teachers do not think that they should be held accountable for students' scores on competency tests. I suspect that this result should be interpreted to mean that student failure is not the teacher's responsibilty, though they might be willing to take some credit for success. Other items on the list either reveal mixed opinions, or else duplicate the points mentioned above.

What factors are correlated with variation in responses to the questionnaire? This part of the analysis in still is the early stages, and the overall plan is not yet clear. Some of the preliminary findings merit comment, however. For instance, the amount of mandated testing in districts that report no minimum competency testing is threetimes as much as in districts with competency tests. This is a striking result, if not artifactual, and deserves further examination and interpretation. The findings also indicate that practitioners who have had experience with centralization of the assessment process do not care for it, but that it easily becomes a way of life (low SES districts, who receive categorial monies in return for an increase in mandated testing, view such assessments as more important than higher SES districts, or so say the principals). Finally, the general pattern of correlations among the variables is weak (only 1 out of 10 correlations is greater than .33, and even here there are probably artifacts due to multicolinearity, a dread disease of correlation tables).

Definitions, Purposes, and Audiences

The data from the survey are still being analyzed, and further clarification of questions raised above both explicitly and implicitly will undoubtedly be forthcoming. Nonetheless, I think that it is possible to comment on two questions posed earlier that are addressed by the project: what operational definitions comprise significant conceptions of testing among practitioners, and what purpose and audiences are served by these definitions?

It appears to me that there are two operational definitions to be found in the data. One definition is primarily rooted in the teacher's judgment, in observations, conferences, teacher-made tests, curriculum-embedded tests, and other sources of evidence that are internally generated (i.e., within the classroom). This definition of assessment is "rich and soggy", subjective, dynamic, and interactive. It is expensive and time-consuming and takes place over the entire course of the school year. The chief criterion is validity for instructional purposes. A second definition springs primarily from external sources: it is the popular conception of the test. The important criteria are objectivity and efficiency.

As to purposes and audiences, it appears to me that there is in place a testing machinery that is now taken for granted--the second definition mentioned above--a machinery of uncertain validity, but one that serves the purposes of evaluation of school achievement for administrative and legislative audiences, and that is used to inform the public about the state of the schools. This purpose is a relatively thin one, but of considerable significance; like the Dow Jones average or the Commerce statistics on the unemployment rate, these measures are important in the shaping of public opinion, and they provide a rough index of whether the situation is getting better or worse in general. More detail is needed if an individual wants to invest some money or needs to find a job.

Teachers do not fully trust the technologial definition alluded to above, probably with good reason, given the purposes that assessment must serve for them. Unfortunately, it appears that they do not possess a clear-cut alternative conception that they hold with any confidence. If this conclusion is correct (and it is a reading that, while consistent with the survey data, is nonetheless not forced upon the reader), and if you think that assessment should be the handmaiden of the curriculum, then you might well wind up thinking that we have a serious problem before us.

Bank and Williams (1981) note that schools lack a "technical core"; unlike the professions of medicine and law, education cannot point to cornerstones of clearcut substance such as biological science or precedent. These writers propose that we fill the gap by the new technologies of criterion-referenced tests and "identifiable teacher behaviors" that have been empirically correlated with test performance. This proposal is set forth as a solution to the testing-teaching link (p. 52).

The proposal is an intriguing one. It effectively does away with the conflict between the two definitions presented above, by deleting the "soft" definition based on teacher judgment. The ultimate payoff from this approach depends on the degree to which present-day test design is adequately matched to a valid representation of the curricula of the school. If the curricula are to be defined as "whatever tests measure", then this appraoch is quite satisfactory. However, some of us suspect that there is a curricular reality that stands apart from the technology of testing, and that is the ultimate validating criterion.

Let me express my position quite directly. The CSE test use survey does not entertain the possibility that the research question as operationalized may be off the mark (remember the story of the drunk looking for his wallet under the streetlight). Teachers may have good reason to ignore the bulk of the test information that is provided to them, if this information lacks validity for the task that confronts them: instructing students.

It might be more enlightening to search for ways to aid teachers in becoming more articulate and confident in their conceptions of the role of assessment in relation to curriculum and instruction. It might be more worthwhile to search for ways to help teachers to refine their uses of observation, work samples, teacher-made tests, and professional judgment. It might be more appropriate to develop techniques for bringing these kinds of data into the system. The district and the public might be better informed if assessment were grounded in the teacher's professional judgment rather than the results from multiple-choice instruments.

CSE has broken important ground in its exploration of test usage by teachers, and in seeking to lay bare the perceptions of teachers and principals about the meaning of these activities. Further analysis of the survey findings is called for, and we must hope that these results will be forthcoming in timely fashion. The data, while limited in some ways, address one of the most important issues for evaluation of the work of the public schools, and it is vital that the message in the "runes" be examined in detail and with care. And there is clearly more work that needs to be done.

References

Bank, A., & Williams, R. Annual report of the Evalution Design Project to the National Institute of Education. UCLA Center for the Study of Evaluation, 1981.

Dorr-Bremme, D., Choppin, B., & Burry, J. Annual Report of the Test Use Project to the National Institute of Education. ULCA Center for the Study of Evaluation, 1981.

Dorr-Bremme, D., & Herman, J.L. Testing in the schools: A national profile. Paper presented at the 1983 summer invitational conference.

Testing In the Schools: An Ethnographic Perspective

Harry F. Wolcott
College of Education
University of Oregon

Some time ago in casual reading I came upon one of those good sentences in a reviewer's quote that compels one to take, and make, note because it either challenges, echoes, or helps to clarify one's own thoughts on an issue. The issue in this case is powerlessness. The sentence, taken from Nancy Henley's long-titled Body Politics, Power, Sex, and Non-verbal Communication, is itself short and straightforward: "The power of disruption is the ultimate power of the powerless" (Henley, 1977, p. 83). I will return to this sentence after some introductory comments.

The interest in "Testing in the Schools" expressed in this summer conference, invited responses to a specially prepared paper (Dorr-bremme & Herman, 1983), and an ambitious long-term study conducted under the auspices of UCLA's Center for the Study of Evaluation, provide the focus of our coming together in a common enterprise, further testimony to the important role that testing has come to assume in American education. But my own role in this endeavor is not all that clear. I am not known for my contribution to the study of testing. If I am thought to be a contributor to the field of educational evaluation, it must be by people who have not heard me rant and rave on behalf of using ethnography as an alternative to evaluation rather than as an alternative way of doing evaluation (see, for example, Wolcott, 1975, 1982a, 1982b).

As for tests themselves, I have never enjoyed taking them and have seldom performed superbly on them. (It used to be at such tasks as calculus, German, organic chemistry, and the Reader's Digest concern for my word power; today it's blood pressure and eye exams, but I still don't seem to do superbly or to improve significantly without extra help. It took special study sessions then; it requires pills and bifocals now--and trifocals are on the way, so that I'll see what my eye doctor wants me to see. He can't stand to have me below average in "seeing" if I am going to remain one of his patients.)

Based on personal and not particularly pleasant experience as a youth subjected to years of involuntarily being tested, once I became a teacher and thus a potential tester-of-others I never used machine-scored exams and have rarely used multiple choice tests. Usually I give no formal exams at all. If I do, exam grades are subordinated to grades on brief papers and, especially, on term projects. University students whose only talent is at test-taking probably avoid my classes. On the other hand, students occasionally thank me for making them organize and present their own ideas carefully and for trying to help them write better under circumstances where what-one-has-understood takes precedence over speed, short-term memory, and intelligent guessing!

Of course, I realize that in my eagerness to editorialize about test-taking I skipped too quickly over the term "ethnography". Regardless of my efforts within the field of education to keep ethnography separate from evaluation--primarily so that evaluators become grist for our mill rather than ethnographers becoming grist for theirs--the invitation for me to discuss issues of testing and to comment on CSE's project must be related to my interests in ethnography, the descriptive and interpretive research approach of the cultural anthropologist doing fieldwork. I have written and spoken on this topic often. At present I am preparing an invitational paper on ethnographic research for a book being edited by Richard Jaeger and sponsored by the American Educational Research Association that will present a number of these so-called "alternative approaches" to educational research.

Yet I feel a certain sense of caution in attempting to offer some "ethnographic perspective" on testing in general or on CSE's research efforts in particular. I am not all that familiar with the project, its contractual obligations, or the particular interests of its research staff. Nor do I know at this point how far-ranging one can be with a project that is essentially complete. But of this I am certain: any further ethnographic explorations can only expand the complexity of the project's scope or findings in looking at the numerous ways tests and testing may be used and abused by people who make them, give them, take them, or interpret them.

Further, I do not have an adequate sense of what CSE may already know that would be of interest to me as an ethnographer. That is because, to my ethnographic dismay, the summary paper to which I am responding here was developed around survey results and alludes only occasionally to interview data. That is, of course, a standard and acceptable practice in educational research, but it is exactly the opposite of the way I would proceed if I were doing and reporting the research ethnographically.

From an ethnographic perspective, had I been preparing the report, I probably would have presented a number of well-contextualized case studies, providing instances (stories, if you like) of formal and informal testing in their "natural" classroom setting. Perhaps my cases would be derived from teacher interviews, perhaps from interviews with students; hopefully they would reflect both what people told me and what I observed them doing over extended opportunitites for observation. Perhaps my case examples would compare testing at a few different schools or in different classrooms in the same school, but the case studies definitely would provide opportunity for my readers to get to know some individuals, or some settings, very well. Project requirements allowing, I might provide all the case data from only one classroom setting. Later, with adequate survey data, I would then try to place the case or cases as somewhat typical, or atypical, or characteristic of certain conditions but not of others. Ethnographers don't worry about finding "typical" cases; they worry about adequately specifying or "locating" the cases they present. In contemplating human social life, concepts such as "typical" and "average" must be regarded with great caution.

In the paper prepared for us, the survey results provide the promised "profile" that CSE sought to obtain, but it is a profile of everyone--and thus of no one. I suspect that CSE has a great deal more data--as yet unreported--that address an issue one might probe in greater depth: in what ways do teachers use tests devised by others so that they retain their own sense of power over their classrooms, and how does their individual understanding of tests correlate with their tendency to use tests in the way that test authorities--the "high priests" of testing--say they should be used? I can even point to a kind of hypothesis: the more that teachers understand about classroom tests prepared by someone other than themselves, the more they will behave appropriately "test-wise" and the less they will exhibit behaviors toward testing, tests, and test results that are irrelevant, irrational, inconsequential, or at least inappropriate. In its adverse, the hypothesis becomes more interesting: teachers who do not have a clear idea of what formal tests can and cannot do, how best to use them, or what the limits are on their findings, will use more strategies for dismissing test results, or finding fault, or becoming defensive, or making exceptions, or confounding results by the way they prepare their students for the tests or conduct the testing.

My point is this: testing is powerful business. Testing is variously perceived and variously understood by classroom teachers. CSE's survey research indicates that the tests that teachers themselves devise are the ones they rely on in the daily course of affairs. For some proportion of the total population of classroom teachers that CSE may already have identified statistically, the testing that teachers are required to do as a condition of employment probably increases or supports a prevailing sense of powerlessness. And that brings me to that sentence I quoted at the outset: "The power of disruption is the ultimate power of the powerless" (Henley, 1977, p. 83).

I do not recall the context that prompted author Henley's observation--most likely it dealt with totally institutionalized people, not the partially institutionalized population of the schools. But I am becoming increasingly intrigued by the myriad ways we humans devise, and regularly employ, for disrupting the systems and institutions that seem forever on the verge of overwhelming, dehumanizing, or otherwise consuming us. Seemingly powerful, even ruthless schemes and organizations face a formidable foe in the responses they provoke from their human constituencies. We cope with what we perceive to be wrong or misguided in the goings-on about us by constant disruption, non-compliance, re-interpretation, and so forth. Somewhat euphemistically such behaviors can be numbered among the "adaptive strategies" that humans devise for coping with the world about them. From the individual's point of view, disruption is an effective, and at times a quite personally satisfying, adaptive strategy.

As time, effort, and interest allow, in the present study or in future ones, I encourage CSE and others to look more closely at the full range of adaptive strategies teachers employ to cope with imposed classroom testing. I realize that I've approached this topic with my anti-test bias showing--because, for one thing, I think we greatly overdo testing, gathering far more data than we ever intend to use, and, for another, because we so often use testing to shift blame onto test takers. Nevertheless, inviting attention to the range of adaptive strategies that teachers use does not require one to load the dice for or against such behaviors, it merely calls attention to looking closely at the uses people actually make of what is available to them, be it freely chosen or dogmatically imposed.

Specifically, here are some questions that would interest me were I working with CSE on this study of testing in the school:

1. What do teachers themselves include in the full range of activities they consider as constituting their classroom testing? (But ask them, don't tell them, as was done on the questionnaire. Let them talk; develop the categories later.

And consider the possibility that teacher concerns are for assessment, broadly conceived, rather than with testing, narrowly conceived.)

2. What are the ways that teachers use tests in classrooms? (How long a list would it take to decribe all the reasons teachers have for testing?)
3. What do teachers understand about testing itself? What is their "knowledge base"? Why have they learned what they have learned, and under what circumstances might they be interested in knowing more?
4. What do teachers actually do with test information gained from tests devised by others:
   a. when they want it, seek it, and agree with the results?
   b. when they don't want it, seek it, or agree with the results?

With CSE's own long collective personal and professional experiences with testing, the new survey data, and, especially, the new and as yet largely untapped bank of interview data, I believe CSE already has much to say on these issues. In examining the complexity of the ways teachers use tests, I think we have an opportunity to learn about the practice of teaching as well as the practice of testing. In looking at testing, there is also the opportunity to explore more of the uncertainty associated with teaching--the risks to which teachers feel exposed and how they protect themselves from such risks.

The ethnographer in me would insist that a few cases well studied would enable me to begin to understand the ways that teachers really use tests in class. I am sure CSE can also design effective ways to assess what teachers know about tests, and I would find that information--reported in specific detail--very informative.

As a somewhat critical aside, I'm not sure that CSE's survey research has made any breakthrough in measuring attitudes or in demonstrating that attitudes tell us very much. Let me suggest how CSE has succumbed to "standard testing procedures" in devising the "instruments" used, and how both the "testing" procedures and the responses of the "test-takers" are also part of the whole context of testing that an ethnographer would want to examine. On a rigorous 4-point scale ranging from the crisply clear phrase "Strongly Agree" to the equally crisp "Strongly Disagree", respondents in the carefully-structured sample were instructed to indicate their "level of agreement" to such crisply worded statements as:

1. "Commercial tests are usually of high quality", or
2. "The pressure that testing exerts on the schools has a generally beneficial effect", or
3. "Tests of minimum competency are frequently unfair to particular students."

These are CSE's questions, I've only added the emphasis. CSE has acquired the sub-culture of educational test-makes well. (I must note that when I see questionnaire items like these, I never feel very defensive about the criticisms aimed at the "softness" of ethnographic research.) Consider the assumptions of the second of the three items in my not-so-randomly-selected list, the question about beneficial effects. To answer the question at all, one has to accept as fact that testing exerts pressure on schools. I do happen to feel that testing does exert pressure on teachers and on schools, but as a teacher-respondent to a survey like this, I'd rather tell CSE than have CSE tell me. Consider further that against the crisply clear phrase "generally benificial" I must discern the nuance between "Agree" and "Strongly Agree". The use of "generally" diminishes the power of my selecting between "Agree" and "Strongly Agree". One way to handle that kind of invitation to powerlessness is by not completing the questionnaire. Ther silent majority of 40-50% non-respondents have disrupted CSE's study and diminised the results simply by doing nothing! That also raises questions about the seriousness with which dutiful respondents may have completed CSE's interrogation of their attitudes.

Against the potential ambiguity in survey questions such as these--and clearly I am taking the survey instrument as yet another form of "testing in the schools"--I'm far more impressed with what I can learn from even a brief quote like the following, taken from one of CSE's own interviews. At first blush this statement appears ambiguous or at least low-key, but in fact the ambiguity can be (and in this case was) interpreted as conveying some tentativeness that maintains control in the hands of the teachers. For whatever type of test is being described (and I think we need that level of specificity, although it is not provided in the excerpt), this particular teacher does not reveal a sense of powerlessness:

> You can't count a score on one test too heavily. The kid could be sick or tired or just not feel up to doing it that day. Maybe his parents had a fight the night before. Maybe he

> doesn't try. Maybe he doesn't test well.
>
> (Dorr-Bremme & Herman, 1983, p. 12)

But wait! while we've got even one teacher on the line, I'd like to know a bit more. Here are six reasons for not counting a score on one test too heavily, but they are for a hypothetical case. What about a real case in that teacher's own classroom? For that one particular teacher, what constitutes enough tests that the scores can no longer be discounted? What about the test scores of kids who always come to school tired? Or of kids who hardly ever come to school? There is lots more that I would like to ask. I'd be willing to trade in most of the survey data for a closer look at a case or two of classroom testing in process, or teacher elation at good test results (one might study only that) or teacher panic at bad results (one might study only that). But, given the data already at hand on this project and the talents of the research staff, maybe we can have both. I urge CSE to make fuller use of interview and observation data in the final report, at least in "fleshing out" the numbers in the comprehensive survey by augmenting them with anecdotal data that suggest what teachers mean by their responses.

Along the way, maybe we can give more thought to the issues of what constitutes "beneficial pressure" in the schools and whether that's the kind of pressure that the current emphasis and reliance on testing now exerts. My hunch is that, like virtually everything else we do in schools, the pressure that testing exerts in schools is beneficial to some teachers and to some students. Testing is probably of greatest benefit to the commercial developers who make and sell tests. Most recently it has also become "beneficial" to school critics in particular and politicans in general.

The fact of life is that, beneficial or not, everyone connected with schools has to cope with testing. An ethnographic question is "How do they?" A focus for so broad a question is the one I have suggested here: a closer look at whatever relationship may exist between any one teacher's sense of powerlessness and that same teacher's capacity for disrupting the would-be orderly world of the test designer. That is a good proposal-size question to address. It may also contribute to our understanding of a larger issue: how did evaluation ever come to occupy such a central place in the ethos of American educators?

References

Dorr-Bremme, D., & Herman, J.L. (1983). *Testing in the schools: A national profile*. Paper presented at the 1983 summer invitational conference of the Center for the Study of Evaluation, UCLA, Los Angeles.

Henley, N. (1977). *Body politics, power, sex, and non-verbal communication*. Englewood Cliffs, NJ: Prentice-Hall.

Wolcott, H.F. (1975). Fieldwork in schools: Where the tradition of deferred judgment meets a subculture obsessed with evaluation. *Anthropology and Education Quarterly*, *6*(1), 17-20.

Wolcott, H.F. (1982a). Differing styles of on-site research, or, if it isn't ethnography, what is it? *Review Journal of Philosophy and Social Science*, *7*(1,2), 154-169.

Wolcott, H.F. (1982b). Mirrors, models, and monitors: Educators' adaptations of the ethnographic innovation. In G.D. Spindler (Ed.), *Doing the ethnography of schooling*. New York: Holt, Rinehart and Winston.

In Tests We Trust?

Remarks On the Pattern of Test Use in Our Schools

Philip W. Jackson
Department of Education
University of Chicago

Participants in this conference have been asked to do the following: 1) identify an important question or area of concern in testing and/or education, 2) discuss the findings of the CSE-sponsored survey in the light of the identified questions, and 3) identify next steps for research and/or policy and practice. I would like to modify those directions only slightly. I will begin by identifying four trends that seem to me to stand out in the data of the survey report. I next will comment, rather speculatively I fear, about tests in general and some of the assumptions that underlie their use in our schools. Finally, I will seek to show how those speculations bear upon the identified trends in the data.

I.

Without in any way intending to be critical of the survey, I think it fair to say that its findings, at least in gross outline, are anything but surprising. What they tell us in general about the use of tests in our schools most of us already know, which is that tests are widely employed by both teachers and administrators; that they are used for a variety of purposes, from decisions having to do with individual students to public relations efforts on the part of the central administration; and that the impact of mandated testing is evident at both the elementary and secondary levels.

In addition to those bland and predictable findings there are several other general trends which, though not quite so predictable perhaps, are also not terribly surprising. Four of those trends strike me as being noteworthy, however, for reasons soon to be explained. The first is that tests are rated as being of greater importance for school decision-making of almost all kinds in schools serving lower SES students, as contrasted with those serving higher SES populations. This tendency is most noticeable in the finding that in 26 of the 32 possible comparisons between higher and lower SES schools, the latter receive the higher mean rating. The only two types of decision-making for which that overall trend does not hold are those having to do with teacher evaluation and student promotion.

The second of the four trends that caught my eye reveals that among the three groups of teachers questioned, high school math teachers appear to be most favorably disposed toward the usefulness of testing, high school English teachers next, and elementary teachers last. This shows up clearly in Table 12 of the survey report. On the two items under the category of "usefulness of testing", the percentages of agreement reflect the progression I have described. It is equally important to note, however, that on the item asking whether tests of minimal competency are frequently unfair to particular students, and on the one asking whether the pressure testing exerts on the school has a generally beneficial effect, the same tendency is clearly evident. Secondary math teachers are least likely to call minimum competency tests unfair and most likely to laud the beneficial effect that the pressure of testing exerts on the schools. The reverse is true for elementary teachers, with secondary English teachers falling somewhere in the middle on both items.

The third result to which I would direct attention shows more time being devoted to testing in high schools than in elementary schools. This is clearly evident in Table 1 of the report. The annual amount of time spent on testing in tenth grade English is almost three times as great as the comparable figure for reading instruction in the elementary school. For mathematics, comparing high school and elementary classes, the difference is twice as great in the same direction. Translated into numbers of testing sessions, the breakdown reveals the average elementary school youngster being tested every couple of weeks in both reading and math, whereas tenth grade English and math students are tested once a week in each of those subjects, and sometimes more often than that.

The fourth and final finding to which I would like to draw attention has to do with the teachers' preference for tests of their own making and, more importantly, for their own observations and opinions over tests of any kind. This tendency stands out in Table 5 of the report, which displays the teachers' ratings of various kinds of tests, including their own observations, as devices for helping them make a broad range of educational decisions. With respect to each and every kind of decision, the ratings of importance rise steadily as we move from standardized test batteries to the teacher's own observations, with district and minimum competency tests, tests included with curricular materials, and teacher-made tests being the three categories lying between those extremes. On a four-point scale of importance, with four being "crucial", only the teacher-made tests and the teacher's own observations consistently receive a mean rating higher than 3.0. It is also worth noting that the principals, too, placed a higher rating on teachers' opinions and recommendations as aids to decision-making than they did to any of the several forms of testing.

So much, then, for the four trends that I find to be notneworthy: a greater reliance on tests in lower SES schools, tolerant attitudes toward tests showing up as being greatest among math teachers, testing taking place more frequently in high schools than in elementary schools, and teachers relying more on their own observations than on tests of any kind. That pattern of test usage and of attitudes toward tests may not be all that surprising, as was suggested at the start, but I find it to be intriguing all the same. To say why requires some talk about tests in general and about their classroom use in particular, a task to which I now turn.

II.

Paper and pencil tests of the kind found in schools are so common that little if anything need be said about their gross characteristics. We all know, for example, that they are usually tests of knowledge or skill of one sort or another, comprised of a series of questions students must answer or tasks they must perform. We further know, as the items in the survey report repeatedly remind us, that tests may be commercially produced, designed by specialists for a single school or school district, or prepared by individual teachers for exclusive use in their own classrooms. We know many other things about tests as well, such as what they look like as physical objects, how to bone up in preparation for taking them, what it feels like to succeed or fail in that task, and so forth. About these and related commonplaces, further comment is unnecessary.

What it is necessary to say something about, however, are the presuppositions that underlie the construction of tests, together with some of the less frequently discussed reasons why tests are seen as very useful, if not indispensable, tools for today's classroom teachers and for others as well. The reason these remarks are necessary is simply that most of us tend not to think about such matters very much, a tendency in need of countering from time to time. Or so it seems to me.

The enabling presuppositions that lie behind the development of the kinds of tests used in our schools today are both epistemological and ontological in character, which is to say they have to do with our ideas about knowledge and its properties, particularly those properties having to do with the essence of knowledge, with how real it is. One way of inquiring into the reality of knowledge is by examining our customary manner of thinking and talking about it. What such an exercise reveals is how knowledge compares with other features of reality, how it resembles other things we call real, like an orange, say, or a stone, or a sack of wheat. For example, we speak of knowledge existing, as we do an orange. We speak of being able to weigh knowledge, as we do a stone. We talk about the spread of knowledge, as we do wheat. How else do we customarily think and speak of it? What further might be said of its "ontic status", a term philosophers themselves sometimes use to speak of the essence of being?

Well, for one thing, knowledge, as popularly conceived, is said to exist in units. It comes in bits and pieces that can be counted and sorted in a variety of ways. The smallest of these, when verbally expressed, is variously called a fact, a proposition, or, more colloquially, a piece of information. When skills are being talked about, rather than verbal or propositional knowledge, the equivalent unit of smallest size is a movement or a physical position of some kind, such as the proper way of gripping a tennis racquet or how to position one's fingers on a typewriter. These rudimintary elements are often referred to as "basics" or "fundamentals". The largest unit of knowledge in common parlance is a "body" of some sort, though terms like "domain" and "field" are also commonly used to refer to macro-units of what is known.

One of the most important properties of knowledge is its truth value. It is also one of the most troublesome when it comes to establishing the existence of knowledge within an educational context. To see why this is so, we need consider very briefly the difference between the outlook of a professional epistemologist and that of a practicing educator.

When the professional epistemologist speaks of something called "the truth condition" as being necessary for the establishment of knowledge, he or she is usually referring to some means, either empirical or logical or both, by which a correspondence of some sort can be established, a correspondence between, say, the world of language on the one hand and the physical world on the the other. Only when such a match can be affirmed, most epistemologists would insist, is it legitimate to speak of genuine knowledge. There is much more to the epistemologist's concern than this, of course, but basically the truth he or she is interested in is of this relatively abstract and formal kind.

The truth about which teachers and other educators are chiefly concerned resembles that of the professional epistemologist in some respects but certainly not all. It has less to do with a formal property of knowledge per se than it does with the very practical questions of whether a given piece of knowledge or perhaps a whole body of it "resides", so to speak, within the student or students to whom it has supposedly been transmitted. The kind of correspondence typically sought by the educator is that between the teacher's knowledge (or that of the textbook) on the one hand, and the student's knowledge on the other.

It has already been pointed out that we commonly think of knowledge as having the property of being disseminated or "spread" the way, say, the contents of a sack of grain might be. This dissemination can take place in a number of ways. Knowledge can be passed along from one person to another or from one person to many others. It can even be passed from books to people, as we well know, and recently other technological inventions, such as television and computers, have come to play a part in the process.

Unfortunately, however, we know equally well that the transmission of knowledge, by whatever means, is not always as successful as planned. Like grain, it does not always lodge and take root as we would like it to. To find out whether it has or has not is where tests come in. Or almost.

The discovery of whether a particular unit of knowledge has been received as sent or directed would seem, almost by definition, to require a deliberate inquiry of some sort. Not always, however, for the simple reason that sometimes the receipt of knowledge is spontaneously registered, as is alleged to have happened when Archimedes gave out his famous cry of "Eureka!" as he ran naked through the streets of Syracuse. Sometimes its reception is more discreetly conveyed, as, for example, by a simple change of expression on a student's face, revealing to one and all that, as the saying goes, "the light has finally dawned."

More frequently, however, some kind of deliberate action does have to be taken if we want to find out whether or not someone knows something. Three kinds of action are common. Others might be as well, but these three are so logically compelling that they come to mind at once.

We first might patiently wait around for the knowledge to be naturally expressed or acted upon. This non-intrusive approach has the obvious advantage of assuring us that the knowledge we are interested in is not only possessed by the person or persons to whom it has been transmitted, it is also being put to use by them. The obvious disadvantage of this approach is that for most kinds of knowledge it would simply take too long to find out what we want to know. Also, it is easy to see how such an approach might well entail ethical problems of no small consequence. Following a person around, while waiting for him or her to display the knowledge we are looking for, may not always be the most welcome form of companionship, to say the least!

For these and other reasons, this most "natural" and "non-obtrusive" method is seldom employed in educational settings. About the closest we come to it within schools is in instruction in athletics and the performing arts, where a coach teaches a particular skill and then sends his or her students into the fray, so to speak, while he or she watches them from the sidelines or from offstage in hopes of discovering how well their lessons were learned. Teachers of other subjects may keep their eyes out for such naturally occurring signs of knowledge acquisiton as well but, aside from the exceptions named, few if any rely exclusively or even heavily on this evaluative technique.

A second general strategy, as natural in its own way as the first but of a different kind entirely, is simply to ask the would-be possessor of knowledge whether he or she knows whatever it is that is being taught. "Do you know X or don't you?" the teacher innocently inquires. This happens all the time in classrooms. We see it most clearly when teachers ask students to reveal their understanding of something by a show of hands or a nod of their heads. The crucial point of this familiar practice is that the teacher's query stops with the answer to the questions asked. He or she accepts the students' testimony as received and moves on from there.

A closely related practice within schools in general is to seek documentary evidence of knowledge acquisition, such as a transcript, a letter of recommendation, or a diploma of some kind. Here, too, what is being relied upon is testimony of a sort, rather than the actual display of knowedge. These documentry procedures are common in admissions offices, where great reliance is placed upon official statements of one kind or another having to do with what prospective students allegedly know. The same is true of school officials whose job it is to determine a student's eligibility for promotion or graduation. Both sets of decisions almost invariably make use of test scores as well, but note that as far as the admissions officer or the dean of stduents is concerned, such scores function in very much the same way as do nods that signal understanding to the teacher in the classroom. The scores themselves are evidence of knowledge, more reliable than personal testimony perhaps, but at least one step removed from its direct revelation, as are nodding heads and raised hands.

A third strategy, the one of chief interest to participants at this conference, is to give a test of some kind, requiring that the knowledge in question (usually just a sampling of it) be actually displayed. This is commonly done, as we all know, by asking the student one or more questions whose answers, if clearly expressed and if given correctly, reveal the knowledge directly. The equivalent procedure in the case of motor skills is to require the student to perform this or that task.

But the possibility of testing reveals nothing about the necessity for doing so, which is crucial to understanding the place of tests in our schools. So the key question becomes: Why test? Why not simply rely on a person's testimony about what he or she knows or does not know?

There are two answers to that question, each in a class of its own. The first has to do with what we can possibly say about what we know. The second deals with reasons why a person might not wish to give an accurate report of the state of his or her knowledge, even if able to do so.

Some of the limits to our speaking about what we know are obvious. For these reasons alone (and others of the same type could easily be added), the employment of tests of one kind or another becomes a near necessity of many educational settings.

The second class of answers to the question "Why test?" has to do with the harsh fact that what a person says about the state of his or her knowledge cannot always be trusted, not because of limits on what we can properly say about what we know, but for moral reasons. To put it bluntly, he or she might be lying, claiming to know what is not known. There are several reasons why this suspicion is often justified, not the least of which is that very real and unpleasant consequences are often attached to a confession of ignorance in educational settings. Courses may have to be taken again, grades may have to be repeated, report card marks may be lowered, more homework might be piled on, notes may be written home to parents, and more.

Added to the possibility of official sanctions of one kind or another is the social embarrassment that often accompanies the admission of not knowing something. The risk one runs by exposing one's ignorance extends to being considerd "thick" or "stupid" by classmates and perhaps even by loved ones as well. That is not always true, of course. Ask me casually for an item of information that I just happen not to know, and I have no trouble at all confessing my ignorance. But ask me to display what I was supposedly taught, and my inability to do so creates discomfort. Let the questioner be my teacher and let the time between instruction and the teacher's question be extremely brief, and the discomfort peaks. To admit to not knowing what has been specifically taught is, in many situations, much more than a confession of ignorance, it is also an admission of failure.

So there are a host of reasons why teachers might find it desirable and even necessary to administer tests to their students, as opposed to relying upon more informal and casual procedures, to confirm the relative success of their teaching endeavors. Almost any combination of them would suffice to justify the presence of tests in our schools. At the same time, it is important to note that one important sub-set of the teacher's reasons for giving tests, those having to do with the possibility that students might not tell the truth if asked directly, introduces an element of distrust into the whole procedure that, once acknowledged, is hard to disavow. That distrust is intensified and made harder to ignore by the elaborate precautions commonly taken to ensure against cheating during the testing process itself. Students are separated by empty seats, test monitors patrol the aisles, all books and materials must be placed under seats, and so forth. So it is not just that students might be tempted to lie when asked if they knew or understood something. That temptation, our cautionary procedures make clear, carries over into the testing situation itself where it takes the form of wanting to cheat, which is simply another form of lying. Again, not every student feels that temptation, we would hope, and among those who do, not all struggle with it to the same degree. But the temptation is there, all the same, as every seasoned teacher knows. Tests, made necessary in part by an understandable penchant to lie about what we know, introduce for many students an additonal temptation to be dishonest, one in which the consequences of being caught are uncommonly dire.

By calling attention to the human weakness that helps to make tests necessary in the first place, and by pointing to the fact that tests themselves may exacerbate that weakness, making stronger the temptation to cheat and lie, I have no wish to condemn the practice of testing in general nor to speak out against the use of any particular kind of tests in our schools. On the contrary, when it comes to weighing the pros and cons of testing, I believe the stronger argument to be on the side of tests and all they have contributed to our schools. Their good points are many. Tests have helped in the early detection of learning difficulties. They have contributed to the elimination of certain forms of favoritism and prejudice. They have served to objectify a wide range of educational decision-making, from classroom practices to federal and state policies. Indeed, it is difficult to imagine a system of mass education without the standardization and regularization of that system made possible by the widespread use of tests.

At the same time, it is important to keep in mind the limits and the drawbacks associated with the use of tests in our schools. One of these, I have tried to make clear, is the suggestion of mistrust embedded, so to speak, within tests of almost all kinds and capable of being communicated to the person being tested. That undesirable interpretation may never come across to each and every youthful test-taker, true enough, and there certainly are ways of lightening its impact when it does (just as there are ways of making it more severe if we are not careful), but the danger that it will be so taken remains all the same.

Moreover, it is not only the person being tested at whom the suspicion embedded in tests might be directed. The targets may sometimes include teachers and other school officials as well. The recent insistence on minimal competency testing is a good case in point. Why are such tests being insisted upon? At least a part of the answer has to be that the public no longer trusts the schools to do what they claim to be doing. The public wants proof of the kind that only tests will give. That desire, no matter how politely conveyed, contains the implicit suspicion that all is not well. It is an expression of distrust writ large. Anyone failing to perceive it as such overlooks the core of its message.

A second category of limitations associated with the use of tests in our schools has to do with the restriction of educational aims and goals to those that conform to the epistemological assumptions already mentioned. The most extreme manifestation of these limits occurs when teachers, as the saying goes, teach for the test (the test in questio being established by some external authority over which the tecaher has no control) and do nothing beyond that. Such situations, we would hope, are extremely rare, but they do happen all the same, and when they do the finger of blame cannot be pointed solely at the teachers who accept such a narrow definition of their task. Unenlightened administrative practices, political pressures, and the public clamor for "hard evidence" that has already been cited, all play a part in forcing some teachers to knuckle under to demands that they otherwise would reject.

But teaching for the test is not the only way in which tests might have a constricting effect on the range of educational goals and objectives. A less extreme form of the same phenomenon shows up whenever teachers restrict their efforts to the transmission of "testable" knowledge, ignoring or leaving to others the development of interests, attitudes, values, character traits, and other prized qualities that the schools have traditionally sought to develop.

There is nothing about tests per se that makes such a restriction necessary. That much must be granted at the start. But tests, by their very nature, which is to say by their apparent objectivity and precision and the definitions of the results they provide, make educational goals that are untestable also seem less desirable somehow. What can only be called the "authority" of tests makes seem quaint and old-fashioned, if not downright sentimental, a teacher's desire to awaken his students' interest in a subject, or communicate by his own actions what it means to be intellectually honest, or to show, by daily example, as Socrates did, how puzzlement and wonder can become a way of life.

How real is the danger that the presence of tests will ultimately bring this state of affairs about? How many teachers today actually restrict themselves to teaching what is testable? I confess to having no idea of the number who do, though I suspect it is not very great. Most teachers of my acquaintance, including those who teach subjects that lend themselves to the frequent use of tests, know full well that there are dimensions of their work that elude capture by tests of even the most ingenious design. Maybe I just happen to be lucky in the people I run across as teachers, but I doubt it.

So the danger of most or even many teachers turning out to be Gradgrinds seems, from my perspective, to be rather remote. What is needed, however, to keep their number small (assuming it already is) is the continual affirmation of those dimensions of teaching about which test-makers, by the very nature of their interest and their work, seldom, if ever, speak.

There is much more to be said about what tests can't do, and about the place of the untestable in educational affairs, but this is neither the time nor the place to say it. Let it suffice to insist that tests do have limits, and to warn against having those limits either distort or constrain the mission of our schools. If the public does not perceive that danger, then we must educate them. If those of us who call ourselves educators do not perceive it, then we must be educated as well. All of which brings me back to the pattern of test use with which I began.

III.

To review, the four trends mentioned at the start were the following: greater reliance on test use in lower SES schools, greatest tolerance of tests by math teachers, least by elementary teachers, more testing in high schools than in elementary schools, and greater reliance by teachers on their own observations than on tests of nay kind.

How shall we understand those findings in the light of what has been said about tests and their limitations? What further questions do they raise?

On the surface, at least, they seem far from surprising, as was acknowledged at the start. We would expect tests to be relied upon more heavily in lower, as opposed to higher, SES schools, for the simple reason that that's where concern about teaching the basics is the greatest, and it is in the assessment of that kind of knowledge that tests are most useful. We would expect high school math teachers to look more kindly on tests than do elementary teachers, or even high school English teachers, for the simple reason that mathematics lends itself to testing in a way that most other school subjects do not. We would expect tests to be more widely used in high schools than in elementary schools, for the simple reason that high school teachers typically teach four or five times as many students as do elementary teachers. Keeping track of the progress of that many students is something that tests can help them to do. We would expect teachers to rely more heavily on their own judgments than on tests of any kind for the simple reason that seeing is believing, even when it comes to that kind of esoteric "seeing" involved in estimating how well someone knows something, or what kinds of intellectual difficulties they are encountering, or any of a half-dozen other judgments teachers are commonly required to make.

So what's all the fuss about this pattern of test use? There is nothing at all mysterious about it, or so it would seem. At the same time, I can't help but wonder if there might not be more to the pattern than meets the eye. Given what has been said about the element of mistrust embedded in testing practices, might it suggest that children of the poor are more likely to have that sense of mistrust communicated to them than are children of the well-to-do? If being tested is a form of being on trial, such an experience is encountered disproportionately by the least privileged portion of our school population. Should we worry about that? I think we should.

What about the greatest tolerance of tests among teachers of mathematics? Is that simply a function of mathematics being more susceptible to testing, or might the difference have something to do with the fact that math tends to be avoided by all save those who must have it in order to enter a particular profession or those who, as the saying goes, are mathematically inclined? The choice is surely not either/or, but I suspect that not enough attention has been paid to the latter alternative.

In the light of the same set of observations, does the greater frequency of tests in high schools than in elementary schools have anything to do with the well-documented fact that positive attitudes toward school diminish as we move up the grades? Do high school teachers and elementary teachers subscribe to quite different sets of epistemological assumptions? It is common to call high school teachers "subject-centered" and elementary teachers "child-centered", but, if true, what do those differences have to do with the place of tests in the pedagogical armentarium of those two groups of teachers?

Such questions call for two different kinds of follow-up studies to the one already made. We need to know more than we do about how students perceive the tests they encounter in school, and we need to know more than the survey tells us about why teachers choose to use or to avoid using tests. What is needed, in short, is an in-depth study not just of testing practices but of the sub-stratum of attitudes, beliefs, and opinions that provide the rationale for those practices and that the practices themselves engender.

My observations about the limits that tests might place upon educational aims and objectives, together with the observation that all teachers seem to rely on their own judgments and opinions more than they do on tests, leads me to wonder whether most teachers may not be a whole lot smarter than test-makers and others sometimes give them credit for being. Might it be that the average teacher, even the average math teacher, understands full well that the most important outcomes of schooling, even for those who are still mastering the basics, have little or nothing to do with what shows up on tests of any kind? Might their act of self-reliance contain a message that contradicts and possibly overcomes the suspicion latent in the tests they use? What is the content of that message? As seen through a glass darkly, which is to say as dimly perceived in the statistics of the survey report, I read it as saying something like the following: "Trust in oneself and trust in others are the two most important kinds of trust there are. It is the job of the school to convey that message, loud and clear." Could that be what the teachers are trying to tell us? We don't know, of course, but I for one fervently hope so.

Paths to Excellence: Testing and Technology

A Los Angeles Unified School District Perspective

Floraline Stephens

Director, LAUSD Research and Evaluation Branch

This particular study had much significance for me, because it put into a better perspective the amount of testing required of students at the elementary and secondary levels. Of particular interest was the information that less than 20% of testing is or could be controlled by a central office administration, namely, statewide assessment, minimum competency testing, and norm-referenced testing (commercially published tests).

I considered that when we talk about testing, we do not include the preparation time required of teachers before the tests are given and the processing/scoring time required after the tests are given. A tremendous oversight.

I was quite disappointed with the information concerning the use of mandated testing results, because this testing is terribly expensive (in terms of both human and financial resources). However, it was not surprising that teachers still principally rely on their teacher-made tests, their opinions, their observations, and their recommendations to plan the school year, group students, provide remedial or accelerated work, and grade students' report cards, instead of district continuum tests. The school principals agree that this is correct.

I was interested to learn that principals report not using elementary or secondary level tests to evaluate teachers. This is unusual, because elementary level curriculum test results can monitor what is or is not being taught and how much students learned after being taught. My old mastery learning advocacy position feels that all regular students (not mentally retarded) can learn if they are taught under regular conditions, and that the amount of learning is particularly based upon quality teaching.

What is especially significant are all of the things that either do not happen or happen infrequently regarding test use and teacher accountability:

1. establishing test-score goals at secondary and elementary levels
2. curricular decision making, at the secondary level in particular
3. administrative evaluation of the quality of teacher-made tests
4. continuous monitoring of the use of ongoing progress test data
5. regular or routine procedures established by school or district administrators to help teachers update their assessment skills

This all leads to accurate assessment of student learning or achievement, which has bearing upon appropriate decision-making.

Student learning, or student academic achievement, is something that is not mentioned explicitly in this study. I think that all parties who have roles to play in the student learning process--that is, parents, teachers, administrators, and even the students themselves--want stduents to become academic achievers. All of these persons want students to know how to read well and ably comprehend what is read, compute accurately, and write clearly. Whatever the type of test, we must remember that testing only samples levels of student achievement. Everything that is taught cannot be tested because testing would indeed consume much of the school day, week, or year.

The issue of testing is ever-present, because school districts are caught in a dilemma when on one side the news media constantly report how "good" or "bad" a school district is based upon test scores in reading and math. The public buys this notion of assessing a school's excellence or lack of excellence based upon reading and math scores. For example, my office constantly receives calls from real estate agents trying to place executives and other workers from across the nation in neighborhoods where the schools have "high test scores". Their criterion for excellent, good, fair, or poor schools is based upon a norm-referenced test's median percentiles for reading and math. On the other hand, teachers' organizations lump testing into the "paperwork overload controversy". However, when you question teachers about the time testing takes, they are not referring to their teacher-made assessments of students' progress or lack of progress, but to testing that is organized and coordinated through a school district's central office. They have no control over the type of test or the scheduling. Thus, a sense of powerlessness occurs, resulting in feelings of frustration.

Admittedly, in the Los Angeles Unifed School District (LAUSD) during certain periods of time, principally the spring, it appears that students are bombarded with testing. The CSE study reported that at the elementary level, five per cent of total instructional time in reading and math is used for testing in those subjects. At the secondary level, the percentage allocated to testing is much higher: 20% of English class time and 18% of math class time. The teachers' time expenditures are magnified by preparing for testing, preparing for scoring tests and, if necessary, scoring tests by hand. Combining testing time with preparation time in a compressed time period makes the process sometimes overwhelming.

State Testing

In LAUSD, two types of reading and math tests are mandated by the state--competency testing and the California Assessment Program (CAP). The first is very expensive to administer, while the latter has no direct impact on the district's instructional program because the matrix testing construct does not provide individual test scores.

The LAUSD spent considerable amounts of money to develop competency tests to meet the requirements of the district and of the state. Many may ask, "Why go to the expense of developing your own tests?" When you consider the volume of students (30-40,000 per grade) who will be affected by testing results, the district deemed it extremely important that the tests were fair and equitable to the population tested. The district is assessing what is taught in the district, not a "generalized" version of a national curriculum. This is especially important when students reach grade 12 and can be denied their diploma based upon failure to pass any competency test in reading, math, and language. Not only were the development costs expensive, but the state's annual requirement that the district report an unduplicated count of students by ethnic group passing all three tests is also expensive. While this may seem easy, it is not. The logistics of processing test scores for 30-40,000 pupils is horrendous. Therefore, computer professionals now enter the picture. They do not come cheap. We now have to have computer files, system analysts, and programmers to keep the records accurate. Each year that the state adds another grade level to report about, the workload becomes heavier! The only reimbursement for district mandated efforts is a stipend for each failing student who has a parent conference, and summer school for secondary students who did not pass a proficiency test. Although in California the statewide assessment program is not tedious or long, it is viewed by many as an added test burden. It has little utility because there are no individual student scores to determine individual progress or lack of progress. The state now views this as a possible weakness of its testing program and individual student scores have been proposed for future development. In addition, the state's schedule for testing grades 3 and 6 occurs in the spring period during a very crowded testing period.

Another element that increases the strain of testing is the year-round school schedule. Los Angeles, in order to relieve overcrowded schools, uses a year-round school program. What this means is that there are three tracks of students in session while one track is off session. Testing schedules have to accommodate the various tracks. Therefore, year-round schools have longer periods of testing to accomodate all of the students to make sure they have similar amounts of time for instruction prior to testing.

Norm-Referenced Testing

The Federal Government's Chapter I (formerly Title I) guidelines require assessment of academic programs. The district, because it wants to know how well its students perform in comparison to students across the nation, tests students in grades 3, 5 and 8 with norm-referenced tests in reading and math. You may have read in the Los Angeles Times that our fluent English-speaking students have improved in reading and math in grades 5 and 8, reaching or exceeding the 50th percentile in math. In both of these grade levels, over 90% of the students are classified as fluent English speakers.

Although schools are judged by their norm-referenced test scores, elementary teachers are supposed to use the district's criterion-referenced test results as the basis for improving instruction. The Survey of Essential Skills (SES), cooperatively developed with the Southwest Regional Lab, is a series of CRT's for grades 1-6. The SES is also used as a state competency measure for grade 5. Individual student and school summary printouts are produced which report whether or not students have "mastered" the curriculum for a particular grade level. At the beginning of the year, instructional plans are based upon these test data. Student grouping in many instances is based upon SES sores. Specimen tests or unit tests have been developed by central office curriculum personnel to assist teachers in assessing skills as they are taught before the SES is administered in the spring. However, we learn in the study that this may not be 100% correct, since teachers indicated their teacher-made tests had top priority. In order to reduce some of this crowded testing period, especially in Chapter I schools, the SES has been equated to the CTBS, thus eliminating testing Chapter I students with the CTBS in grades 1, 2, 4 and 6. This effort was viewed by teachers as a significant reduction in the time required for mandated testing! This is important when the study indicates that teachers view time used for testing as time lost for instruction, particularly in loweer socio-economic schools (Chapter I schools).

Study Questions

The Los Angeles Unified School District is not unique in the way that they conduct testing and use test information. There is very little disagreement with the findings in the study. However, responding to the study questions in the context of the LAUSD may add to the information already compiled.

How much testing really goes on? A whole lot in a compressed period of time. The logistics of scheduling and processing hundreds of thousands of tests for the various programs described decreases the amount of flexibility for schools and their staffs. This in itself may make school staffs feel that an awful lot of testing goes on for students with statewide assessment, and minimum competency tests, when in essence according the study, the percentage allocated is quite low in contrast to the use of teacher-made tests or textbook tests.

What functions do tests serve in the classroom? Criterion-referenced testing in LAUSD is used to group students and to pinpoint those teaching objectives that need further emphasis or need to be taught. Curriculum alignment tasks have become an integral part of instructional planning in many of our schools. However, many school staff need assistance in order to do this correctly and have more positive results.

How are test results used by teachers and principals? What kinds of tests do principals and teachers trust and rely upon most? Norm-referenced test scores still seem to be regarded as indicators of success. This is due partly to the media and to school superintendents and boards of education who do make judgments of quality based upon a percentile rank. However, teachers, once they begin to understand the sense of criterion-referenced tests, view them as a hardy aid to improve their instructional emphasis.

The study's overall results can be viewed in three ways or by three questions and the responses to these questions:

1. Is the tremendous expenditure of financial and human resources justifiable when you learn that the decision-making which has a direct impact on student achievement is

principally based upon teacher-made tests? With this information, perhaps the money and time should be diverted from statewide and national testing programs into training teachers to make better decisions about students (such as competency, report card grades, promotions, etc.) by improving their ability to develop good quality teacher-made tests.

2. Is it necessary to have federal, statewide, and districtwide testing to answer the same question: are students improving in reading and math? I would suggest combining some of these efforts through equating studies or eliminating duplicate efforts. This would certainly reduce some of the time loss from instruction.
3. Is competency testing cost-effective in relation to the expected outcomes of having students who can read, compute, and write? This question is important because of the tremendous amount of money and time expended. Since 1979, LAUSD students have been tested to ascertain their minimum levels of competency. Instead of using the money to test more students, the money could be used for a follow-up study to see if, indeed, these student graduates are functioning ably in the real world after leaving school.

Testing In the Schools

Francisco Sanchez

Superintendent, Albuquerque School District

As superintendent, what are my concerns about testing?

1. Are our kids learning what we want them to know?
2. Can we use tests to pinpoint which areas of instruction need improvement?
3. Can we use tests to appropriately select kids for specific programs?
4. What do tests tell us about effective and ineffective teachers?
5. Are tests enhancing instruction, or are they getting in the way (i.e., taking too much instructional time)?
6. In these times of difficult public relations for schools, can we use tests to document value gained form the tax dollar?

Several major questions must be addressed in examining any testing program. Some of them are: why the tests are given, what is done with the information, how well the tests actually match what is being taught, and how testing can help students learn more and learn it more effectively.

Testing is only a part of the process of pupil evaluation and is of real value only to the extent that the results can be used to improve instruction and pupil performance in the classroom. The main purpose of a testing program is to provide feedback to students, parents, and teachers for making decisions related to teaching and learning. A secondary but equally important purpose of the testing program is to provide data for program evaluation so that instructional leaders can examine and modify curriculum.

How much testing is going on in our schools? How much testing is really needed? How will we know when we are testing too much or when testing may be actually intruding into instructional time? Do we as educators have all the information we need to:

1. determine special needs of children
2. help students in specific skills
3. determine if students have retained mastery, or
4. know if the program is successful?

The Albuquerque Experience

In Albuquerque we are required to do state-manadated program assessment. The state has mandated a norm-referenced test battery at grades 3, 5 and 8. The state has also mandated the New Mexico High School Proficiency Examination, which is a "functional literacy" test of life skills and includes a writing appraisal which emphasizes writing production. The New Mexico High School Proficiency Examination is given at grade 10, with additional opportunities to pass the test at grades 11 and 12. Passing this test qualifies the student for a diploma endorsement, or "gold seal", at the time of graduation.

The whole issue of state-mandated testing is an interesting one. In the state of New Mexico, the main purpose of state-mandated testing is to demonstrate district program accountability. However, at the local level we have adopted the position that all testing should be used to improve programs. Therefore, we analyze test results and report them in a number of ways to a variety of audiences, all aimed at specific program instruction.

Our locally mandated tests fall into several categories:

1. Federal program evaluation, which depends primarily on a continuous database of results from a norm-refernced test.
2. Locally-developed criterion-referenced testing, which reflects progress in language arts, reading, math, science, and other academic areas.
3. Diagnostic testing, which determines specific needs and contributes to recommendations for special placement.
4. Performance standards testing which, in conjunction with teacher observation, determines if individual students are prepared for the next level of work.
5. High school course-specific tests, which will (in APS) replace general achievement batteries and will be used for individual as well as program assessment. In Albuquerque, pilots of these tests are currently being analyzed for possible inclusion in the APS Comprehensive Testing Plan.
6. Teacher-option testing, which allows teachers to request test materials from the district's Testing Services Center, which is operated as a sub-unit of Instructional Research, Testing, and Evaluation.
7. Teacher-made testing, which provides the cornerstone of any instructional testing program and reflects essential elements of the real curriculum.

Much of the locally mandated testing, while not specifically required by the State Department of Education, is directly related to state requirements. For example, the New Mexico State Basic Skills Plan requires checkpoint measures of student mastery in several major academic areas during elementary and middle school years. Many of the locally mandated tests fulfill this state requirement and were designed to function as checkpoints.

Use of Test Results

Testing in and of itself may be useless--and actually may be a waste of time--unless results are properly interpreted and used to improve instruction.

Over the past three years the Research and Evaluation Department in APS has implemented a plan for making testing relevant to schools and teachers. With a very limited staff, professional District Program Evaluators have been assigned, as a portion of their workload, to serve in a consultant capacity to specific schools and to work with those schools in interpreting testing and assisting in other evaluation needs. A District Coordinator of Testing coordinates the logistics of testing, analysis of results, and production of testing reports for schools and districts, and provides training in workshop settings for professional personnel.

Using Tests to Improve Instruction

Let me give you one example of a process which we feel has been successful in using tests to improve instruction.

One of the major components of our district testing program is a norm-referenced test battery at grades 3, 5 and 8, which is given annually in mid-March. The process follows essentially the following nine steps:

Step 1: Spring Workshops

In December, the District Coordinator of Testing presents workshops for all school administrators and school test representatives. She is assisted by other evaluation staff as necessary.

Information provided at these workshops covers mechanics of test administration, importance of standardized procedures, testing environments, and implications of testing for instruction.

Principals and test representatives take information gained from these workshops and share it at scheduled staff meetings at their schools.

Step 2: Practice Tests

Practice tests are provided by our Research and Evaluation Department for all schools. Schools are encouraged to use practice tests judiciously to help students learn test-taking skills. These tests are written to address skills which are measured by the "real" test, but never to teach the test itself. Practice test items are designed to familiarize students with typical standardized testing formats and procedures.

Step 3: Test Administration

The test administration occurs in mid-March with testing coordinated at the building level by the principal and the school test representative.

Step 4: Visual Scanning of Answer Documents

Answer documents are hand-delivered by school personnel to the Testing Services Center. Here trained personnel visually scan every answer document, removing extraneous marks and insuring that the required personal information is complete.

Step 5: Test Scoring

Answer documents are delivered to the APS Data Services Center for scoring, with the District Coordinator of Testing working closely with data services personnel in all facets of the scoring process. We have found that in-house scoring not only saves considerable money, but gives the district greater freedom to develop reporting formats which meet the needs of our instructional and classroom personnel.

Step 6: Distribution of Results

Following scoring, the District Coordinator of Testing distributes the testing printouts to all schools. District Directors of Instruction are informed of the release of the data, and assist by working with the school in preliminary examination of the printouts.

Step 7: School and District Reports

The District Coordinator of Testing analyzes the test results and prepares reports for each school as well as for the district. These reports are formatted to display technical data relevant to specific academic areas. For example, each school receives a booklet which includes a separate page for each of the major subtests. The separate page includes the percentage of students falling in high, middle, and low ranges; the number of students who score below national p values (percentage correct) in each skill area; and a comparison of the school values to the national norm group on each skill.

School principals are given a form to use with their staffs to stimulate discussion. The information is helpful to schools not only in identifying children who may require further instruction, but also in determining needs for materials or program modification.

Step 8: Public Release of Test Scores

A press conference is scheduled, with all local media invited, to share the district's report of test results.

Press conferences are planned to coincide with the release of the special testing issues of "APS IN ACTION", a pamphlet published by our own Public Information Office. Timing is crucial, as competition is keen in the media industry and it is important that the press conference be held as close to the release of the Sunday newspaper insert as possible. Therefore, press conferences are typically planned for Thursday or Friday, with articles by reporters being released in the Sunday morning edition. At press conferences all test scores are reviewed, with a complete explanation of terminology, technical specifics, and applications to improvement of instruction. Reporters attending the press conferences are given an early release copy of "APS IN ACTION", as well as a district test report and other pertinent information. Press conferences are planned with a 30-minute explanation of the test scores, followed by an open question and answer session. A panel consisting of the Superintendent of Schools, the President of the Board of Education, the Director of Instructional Research, Testing, and Evaluation, and the District Coordinator of Testing, answers questions relating to testing, curriculum, plans for curricular change, explanations regarding why test scores are high or low, technical inquiries, and other matters of concern.

Reports from the press conference are typically seen on local evening television news broadcasts, heard on various radio news reports, and reported in the morning and evening newspaper. In addition, several reporters often request brief interviews with one or more of the panel members. These interviews are aired over radio and television stations, including the local PBS station, or quoted in newspaper articles. By planning the news blitz and releasing the information in this way, the public is inundated with information regarding the testing program for one or two days, followed by the release of the APS publication on Sunday. "APS IN ACTION" is written in an easy-to-read-and-understand interview format for the purpose of clarifying any misconceptions resulting from the previous two days' news articles.

Step 9: Fall Workshop

Workshops to aid in understanding and applying test data to the instructional program are again offered by the District Coordinator of Testing in August and September. These workshops are attended by school administrators and school testing representatives. Information covered includes how to use the testing report, how to use the item analysis, and how to use the Sample Item Document. The Sample Item Document is a compilation of sample items which relate to specific skills measured on the test. Teachers are encouraged to examine the types of items which are especially troublesome to students, and discuss plans for future teaching strategies which will address those skills, especially those in which students need additional help.

This process works--I'll tell you how I know it works. By following these nine steps, our district's test scores have steadily increased for the past several years. You may suspect, because we have taken such pains to coordinate this effort, that our teachers may be guilty of "teaching the test". This is not the case, and we had a chance to verify that "teaching the test" is not the case, when our state mandated a brand new test two years ago. This new test was kept secure, with no copies available to any school personnel in advance of the testing dates. Other districts using this new test, in its first year of utilization, typically experienced declining scores. In APS, our scores continued to increase, in some cases dramatically. We believe this is because we are emphasizing the skills measured by the test, and never the test or test items themselves.

This nine-step process is followed with other major tests, and always involves Research and Evaluation team members.

APS School Liason Plan

The APS School Liason Plan was designed to supply evaluation assistance as well as to provide test interpretation to all schools. District Program Evaluators work with school principals and faculties to help them understand the meaning of test results and to apply results to their own day-to-day classroom instruction. Annual District Goals lend support to these efforts by continuing to emphasize the importance of instruction to the district.

District Program Evaluators are called on frequently throughout the school year with school requests to explain test scores, assist in conducting climate studies and surveys, conduct needs assessment activities, design and conduct evaluation of programs specific to the requesting school, or other research or evaluation services.

The education of children is a highly complex operation. Reducing the measure of success or failure of education to a set of numbers can be not only overly simplistic, but also misleading or even detrimental to improving instruction. Therefore, great care must be given to insuring that a comprehensive testing plan meets the needs of all participants in the educational process. This plan must span all grades, K-12, and must provide necessary and accurate individual and group information in all academic areas. Information must be provided in a format that is easily understood and easily applied to day-to-day classroom instruction.

Over the years in my work, I've heard a few horror stories about testing which make me realize that we have a long way to go in getting the testing issue under control.

For example, a few years ago a testing coordinator in a large district confided that he had discovered that some students in his district were routinely taking as many as five norm-referenced test batteries in one year. In some cases, students were repeating the same test battery as many as five times! This is evidence of a serious lack of communication, over-testing, and a waste of instructional time! In this case, program directors who were working with the various "special" programs were working with and testing students independently of one another, with no central coordination. In an effort to properly diagnose needs of students, many students were losing many valuable hours of classroom instruction.

Another situation of which I'm aware, and which is currently in practice in one state as a result of state legislative action, requires that all students, K-12, be tested every year on a standardized norm-referenced battery of tests. I have real problems with this approach.

Using one specific test battery for every grade level every year cannot help but influence the district's curriculum in a restrictive way. The longer this practice continues, the more closely the curriculum and the test will match, because public pressure to have high test scores will cause a narrowing of the curriculum to the point where only those skills on that specific test are being taught. I feel this represents an unhealthy control of the local curriculum, and gives a great deal of power to the authors of any one text.

In an Information Age, when knowledge is exploding, this focus on any single test causes a dangerous constriction of the local curriculum, forcing all programs to emphasize only those "basics" represented on the test.

Issues of A Comprehensive Testing Program

A comprehensive testing program will include various types of testing for various specified needs. It will be developed with input from professionals in evaluation, school administration, and classroom teaching. "Ownership" of the program, developed only through a plan of involvement of key people in all areas, will occur only when professional staff have participated in the plan.

In the absence of a sense of "ownership", or when administrators or teachers do not understand the purposes and applications of testing, negative attitudes may become a problem, which will block the full and appropriate use of the test as well as the test results.

Administration and teachers must have a clear understanding of what each test is all about: why it is being given, how it will benefit classroom instruction, and how it will help the teacher understand the student's achievement. Therefore, the purpose of all testing must be made clear and must be thoroughly understood by all users. Teachers and students will only appreciate the benefits of testing if the results and the application of those results are fully explained and made practical to them.

Testing is not a panacea, and does not offer simple answers to all our questions. If administrators and teachers do not fully understand the purposes and applcations of testing, they will be prone to regard such testing as intrusive, of no benefit, or even an absolute waste of time.

Tests for Teacher Evaluation

Using tests in conjunction with merit pay for teachers is an idea which is sweeping the country, and one which should be examined carefully. Let's look at the many considerations of such a plan.

Tests are developed to represent a broad cross-section of curricula across the nation, although no nation-wide curriculum actually exists. Tests developed in this fashion may or may not match individual district curricula; rarely would such a test match a district's curricula in every respect.

The tests which we, in our district, are required to give, vary considerably in congruence with curricula, both by subtest and by grade level. For example, the test at the third grade level appears to measure about 75% of our stated math curriculum, with a considerable amount of our math program not being tested or with several items on the test which are not a part of our program at that level.

As students get older, the match between the test and the local curriculum becomes less close. At the eighth grade and older levels, students take elective courses which broaden their experiences even more, but may never be measured on any standardized test.

Let's assume we wish to determine a teacher's effectiveness by using students' test scores. Let's assume we test all students in the fall and again in the spring, to determine achievement growth during the year. First of all, we must be absolutely certain that the test we are using does match the curriculum we are teaching. Also, we must look at many important variables, such as:

1. Were the students present every day?
2. Do all students have similar previous learning rates?
3. Do students in classroom A have the same ability to learn as those in classroom B?
4. How much growth is enough? How much growth should we expect for teachers to be eligible for additional pay?
5. Are external variables intruding into the learning environment, such as interruptions by announcements or visitors, school assemblies, construction or other noisy activities, student nutrition, and even the time of day the teaching and testing take place?
6. Parent support and expectation of the schools and of their own child is important. We frequently find parents who do not have the time or interest to be involved in their child's education. Many students are required to stay home and babysit a younger child when parents are unable to be home. These students who stay home to help parents are missing valuable instructional time. We cannot teach a child who is not in school.

These and other considerations must be addressed before anyone can use a test or even a set of tests to judge a teacher's effectiveness. When a test becomes too important, or when test results are used for inappropriate purposes, district curricula tend to become limited to only those skills which are covered in the test. A major underlying question is: do we as educators want to allow a test, any test, due to pressures such as merit pay, to drive and control what we teach the children in our schools?

Summary

In summary, a comprehensive testing program should be well-planned to span all grades, K-12. A "scope and sequence" of testing in each academic area, outlining the skills to be measured at specific levels, should be an important part of this plan. A comprehensive testing plan will include norm-referenced testing--each given at specifically pre-determined times in the student's educational career, with careful attention to the purpose of the tests and the use of the results.

Results of all testing should be shared with administration, teaching staff, parents, and those students who are old enough to understand their implications (in Albuquerque we advocate explaining test results to students in fifth grade and older).

In an age of accountability, we accept that testing is here to stay. As testing becomes more and more a part of our lives, we recognize a danger that tests may overly affect our lives, and may even control the academic lives of our students, very possibly in inappropriate ways. As professional educators, our moral and ethical responsibility is to be knowledgable about the purposes and the limitations of testing, as well as to move into the future cautiously, examining all ways of making the curriculum appropriate to the lives of our students, who are the future of our nation.

Testing in the Schools: A Statewide Assessment Perspective

Dale Carlson

Director, California Assessment Program

I consider it a privilege to be at the 1983 CSE Summer Conference, and a special privilege to be able to address you on the implications of recent CSE work for state-level policy--especially assessment policy and its role in school reform. I always look forward with keen enthusiasm to the CSE conferences. Of all the ways which we in California have profited from the work of the Center, it may well be that the conferences have had the greatest impact, or at the least the most noticeable impact. I hope that the small amount of time I'm going to spend on these comments will not significantly decrease the probability of such a benefit accruing to each of you as well.

My comments will take the form of 21 points--21, not because that happens to be the product of the number representing unity multiplied by the number for perfection, or because it happens to be the sum of the six types of points I hope to make. Specifically, I will outline one implication, two limitations, three proverbs, four questions, five whereas's, and finally six recommendations. Actually, more implications are laced in the following narrative, but since the link to CSE's study is tenuous at best, I won't associate the study with the guilt of my biases.

One Implication

There is only one unavoidable implication: There must be more statewide assessment! The data are unequivocal: if statewide assessments account for only three percent of the total testing time, which itself at the elementary level is only five percent of the available instructional time, then it clearly follows that such a service must be offered on a grander scale. (This is a joke. It is only a joke. If this had been put forth as a real implication, you would have noticed a more serious and reasonable point of view propagated.) Obviously such a finding--percent of timewise---does not give state personnel a license for unbridled expansionism, although the observation that growth is said to be the only sure sign of life is not wasted on most bureaucrats. It might suggest, however, that from the standpoint of instructional time as a resource, the relative impact of statewide assessments may be profound and profoundly cost-beneficial--maybe. Obviously there are other costs, including those related to local control which we will discuss shortly, but this conference is not the forum for discussing the seemingly infinite virtues of statewide assessment.

Two Limitations

CSE has done us great service by providing information about the prevalence and ecology of testing in American schools; howver, I do feel obligated to mention two limitations of the study, since they pertain to the task of drawing implications for state policy.

The first is a limitation of scope. It was obviously beyond the ambit of the research to study all the various uses of test results, although one might have expected such from the title: "Testing in the Schools: A National Profile". Other uses, for example, policy studies, resource allocation, and public credibility issues are important functions which must be addressed in any comprehensive look at the effects and uses of testing and would undoubtedly have provided more grist for my mill. Actually this is not a limitation of the study, since I believe it is an unprofessional cheap shot to criticize a study for not addressing the idiosyncratic interests of a reviewer.

Secondly, even within the domain of teacher and principal test uses and as amplified with field interviews, a survey is limited in the types of uses that are allowed to merge. One might wonder if a study of the actual use of tests might yield quite different results, i.e., a study which draws conclusions from actual observations of decisions being made or studying the types of information that are used and how they are combined and interpreted to inform decisions.

The basic problem, however, is one of mismatch between the decision-makers and the levels and types of decisions they make on the one hand, and the types of tests and information supplied by tests, on the other hand. To overstate the case, one could ask, "Why ask teachers what they think of various tests or why or how they use them? Who cares?" I submit that tradition (the democracy of the dead) has led us to believe that it is useful to ask teachers these types of questions--questions which are tantamount to asking carpenters how useful hammers are relative to saws or plumblines, or like asking pilots what types of information they use in making critical in-flight decisions--especially when that information ranges in specificity and logical spatial-temporal relationship to the tasks at hand from such information as altitude and direction to overall policy relevant information such as frequency of air crashes with similar craft.

My point simply is that teachers are interested in process more than goals or outcomes and in immediate feedback to guide their "in-flight" decisions. We should not expect test data to be revered by classroom teachers, and we must guard against the temptation to give undue weight to their comments about the value of test information in the overall process of improving instructional programs.

Three Proverbs

The three proverbs (actually two aphorisms and a poem) are obviously filler material, inert ingredients meant to make the other thoughts palatable. Nevertheless, for the sake of symmetry herewith:

1. Writing free verse is like playing tennis with the net down. (Emerson)
2. To lose one parent could be considered a tragedy--losing both begins to look like carelessness. (Wilde)
3. The shortest poem on the history of microbes:

   Adam
   had 'em.

I'm sure there is a relationship of the above to the CSE study, but it is probably best left to the reader to divine or to safely ignore.

Four Questions of the Naked Emperor, or Four Profane Thoughts About A Sacred Cow

In any discussion of the usefulness of test results to improve instruction, a central theme is that of the match between the intents of the instructional program and the content focus of the test instrument. This is often an issue because of the high value we place upon variation and diversity in instructional programs and on the autonomy of those responsible for instruction in selecting appropriate outcomes. The question is: "Does one dare question the virtues of unlimited variation or unlimited freedom to select learning objectives according to the perceived needs of the learner or the preferences and predilections of the educator?"

In California over the last ten years, the Serrano argument for equal funding has posed the question, "Why should two students be offered education programs of substantially different levels of quality, based on different funding levels, strictly because of an accident of birth, i.e., their residence in different school districts?" For many input and process variables, minimum quality standards are agreed upon: teachers must have a certain amount of training; a certain amount of space must be available for all students; textbooks must meet certain criteria; class sizes must not be allowed to go above certain levels. However, when it comes to the actual intentions of instruction, variability is the norm--indeed, it is the value. One doesn't hear the Serrano argument for curriculum parity. But the question could be raised, "Why should two students living in different districts or attending different schools or having two different teachers in the same school, study fundamentally different topics and have considerably different levels of opportunity to learn a given skill or concept?"

What are the assumptions underlying this seldom-questioned state of affairs? I would like to briefly raise the ugly specter of three questionable assumptions and end this section with an observation.

Is it assumed that the specific goals and objectives of an instructional program are not really important in and of themselves? Perhaps it doesn't matter exactly what is studied and learned as long as something is learned. Is this a manifestation of a latter-day mental discipline, or of a wholehearted belief in the centrality of learning how to learn, or how to think? Perhaps it is an act of faith in the ability of the human mind, if you'll pardon the expression, to sort out and transform the little knowledges, understandings, and skills into a truly meaningful whole incorporating the basic eternal truths, regardless of the specific focus of a given instructional program.

Perhaps the assumption is that schools exist primarily for educators. Is it the inalienable right of teachers and school administrators to decide what the young, in their school, are to learn? If so, does this idea rest on the unstated assumption that teachers should not have to teach subject matter which they consider unimportant or with which they feel uncomfortable? Or is it because we believe that if teachers do not feel comfortable with or are not well trained in a given field, they should not confuse the students with poorly presented information and poorly monitored practice, reinforcement, and assessment. Teachers' groups frequently mention the need for a greater role for teachers in curriculum development; however, over the years, teachers have made virtually all the important decisions about what their students learn.

A third general assumption could be that local control is everything: regionalism is paramount. This assumption implies, of course, that we do not live in an era of mass media, instant communication, high speed transportation, and that we are not members of a globally interdependent community. It assumes that citizenship skills and attitudes are substantially different in different areas of the state and country, and that a student growing up in a given community has a greater than chance probability of reaching adulthood in that locale.

Finally, it is a self-deceiving value on variability that we hold anyway, since the heavy reliance upon textbooks, the relative uniformity among textbooks, and the dominance of relatively few textbooks in a given content field means that, de facto, we have a relatively uniform curriculum. The tragedy is not that it is uniform; the tragedy is that it comes about without benefit of democratic dialogue, widespread input, and accepted consensus-forming procedures. The uniformity stems from the preferences of textbook authors as they strive to please editors who hope that they have accurately perceived the latest fads and trends of the marketplace. The final act of the tragedy is that test publishers, in some measure, focus their tests on the content of the instructional materials.

No, I do not believe that we need a lock-step standardized, uniform, centrally-promulgated curriculum, complete with federal inspectors, but I did find it amusing to pursue these interesting strawmen (or maybe not completely strawmen.*) Moreover, it is useful to occasionally examine the nature of our educational values. (I have skirted the real reasons that we value diversity, but the purpose of these comments is to soften your thinking to more readily accept some of the principles in the next section.)

---

*Is it encouraging to note that most of the recent national studies of quality and reform of American education call for more assurance that courses with the same name share a certain commonality of content emphases?

Five Whereases

The five whereases are foundational to the six recommendations to follow. The five whereases are articles of faith which I hope you are willing to grant me in order to be able to present the recommendations.

WHEREAS THE CRISIS IN AMERICAN EDUCATION IS REAL. The crisis is, of course, not only one of intellectual dimensions, but we will focus on that aspect for this line of reasoning.

WHEREAS A SET OF COMMON CORE SKILLS AND UNDERSTANDINGS EXISTS WHICH IS ESSENTIAL FOR ALL STUDENTS TO LEARN TO FUNCTON IN AND CONTRIBUTE TO A DEMOCRATIC SOCIETY. Obviously, there are other skills, knowledges and competencies which are unique to a locale and to specific students and subcultures.

WHEREAS INFORMATION ABOUT THE LEVEL OF COMPETENCY OF STUDENTS AT VARIOUS POINTS ON THEIR PATH TO EXCELLENCE IS USEFUL IN HELPING US EVALUATE AND IMPROVE INSTRUCTIONAL PROGRAMS.

WHEREAS TESTS AND SIMILAR DEVICES ARE ONE IMPORTANT SOURCE OF THAT INFORMATION. They are not the only source of this information, and for some goals are definitely not the best source. To paraphrase

E.E. Cummings, "As long as we have lips and voices, lips to kiss with and voices to sing with, who cares if some one-eyed son-of-a-bitch comes along and invents an instrument to measure spring with." Recent floods on the Colorado River, however, indicate the value of measuring and monitoring some aspects of nature, even of springtime.

The Committee on Ability Testing discusses the alleged ambivalence with which tests have come to be viewed in our society. People are allegedly skeptical about the quality and usefulness of achievement tests, and simultaneously skeptical of the quality of our schools, basing this skepticism, at least partially, upon evidence from those achievement tests. I think it could be argued, however, that it is not a case of ambivalence, but a conflict of views held by different groups. I see one group, primarily educators, skeptical of tests; another group, primarily non-educators, skeptical of schools; and a third group, the professional critics, skeptical of both.

WHEREAS TESTS NEED TO BE MATCHED IN LEVEL AND SPECIFICITY TO THE DECISIONS THEY ARE DESIGNED TO INFORM. It seems almost too obvious to be necessary to mention that the level and specificity of the tasks which tests are designed to assess, and the degree to which the information they provide can serve as an indication of performance on other tasks or general cognitive skills, must be different for tests with different purposes.

Finally, we have reached the six recommendations or implications.

Six Recommendations

Recommendation 1: Junior high education must not be ignored.

I think it is mildly significant that the CSE study focused on upper elementary and high school. I think it was a wise decision to do so, yet it is indicative of a general trend to ignore junior highs because of our general long-standing ambivalence about junior high school programs. Nevertheless, any serious attempt to improve American high schools must deal with the junior high issue less obliquely.

Recommendation 2: Broader testing focus.

The focus of achievement testing in America must be broadened beyond the basic skills to include other content areas, such as science and social studies. Consistent with this broader focus would be more intentional effort to focus on the "higher level" problem solving and critical thinking skills central to a real understanding of the nature of these fields of knowledge. Such a move, which, of course, is happening, will not only right the imbalance in the curriculum and the ways in which the tests have been driving the curriculum but, in fact, will allow a greater opportunity for students to better develop their "basic skills" by using them in the content fields. Task structure analysis, information processing, and other tools of cognitive science will be especially useful in mapping out the relationship between instruction and assessment in the area of thinking in the content areas.

Recommendation 3: More vertical integration.

We must get on with the task of designing linkages among local, state, and national (and international) levels of assessment. The National Committee on Excellence called for a national (but not federal) testing program with specific purposes. The advantages in making comparisons with truly representative up-to-date norms and the power and flexibility made available by calibrated item banks are only two reasons why this is imperative. It is a realizable dream. Many of us are, of course, pleased with the expressed intentions of ETS in its winning NAEP proposal to push back the frontiers in this part of the assessment wilderness.

Recommendation 4: Speedier applications of technology.

We need to get on with the clear agenda of refining and exploiting the power of technology to solve vexing testing problems. Tailored testing, for example, and all it represents, is at our doorstep. This is not to say that the problems, and therefore the solutions, lie solely in the realms of hardware, software, and psychometric methodology. In this area of adaptive testing, for example, Bob Wood points out that we need to study the differential effects on student motivation when presented tasks only at optimal difficulty levels.

Recommendation 5: Dual Foci--A call for greater attention to critical distinctions and purposes of tests.

There are several distinctions which need to be fastidiously observed in the design and use of achievement tests to improve instruction. The first of these pertains to the traditional individual-group dichotomy. Historically, our thinking has fixated on aggregation as the key variable; group results were merely the sum of individual scores and, therefore, probably less useful than the results for individual students. The point is made here that group results should be thought of as important data in their own right with unique purposes, not bound to their traditional origin as the sum of individual scores. One might guess that this is a poorly disguised pitch for the virtues of matrix sampling--and one might be right.

There is a second distinction which is mentioned at the risk of falling into Benchley's postulate that people in the world are divided into two groups: those who divide things into halves and those who don't. This distinction focuses upon the object of our interest in the assessment process, that is, either the skill of the person or the skill of the person. Obviously, any piece of assessment data is the result of a person interacting with a task requirement. But one can focus on either the person or the skill. It is argued that the optimal use of test results in the improvement of instruction comes about only with attention to these distnctions both in the design of assessment instruments and in their interpretation.

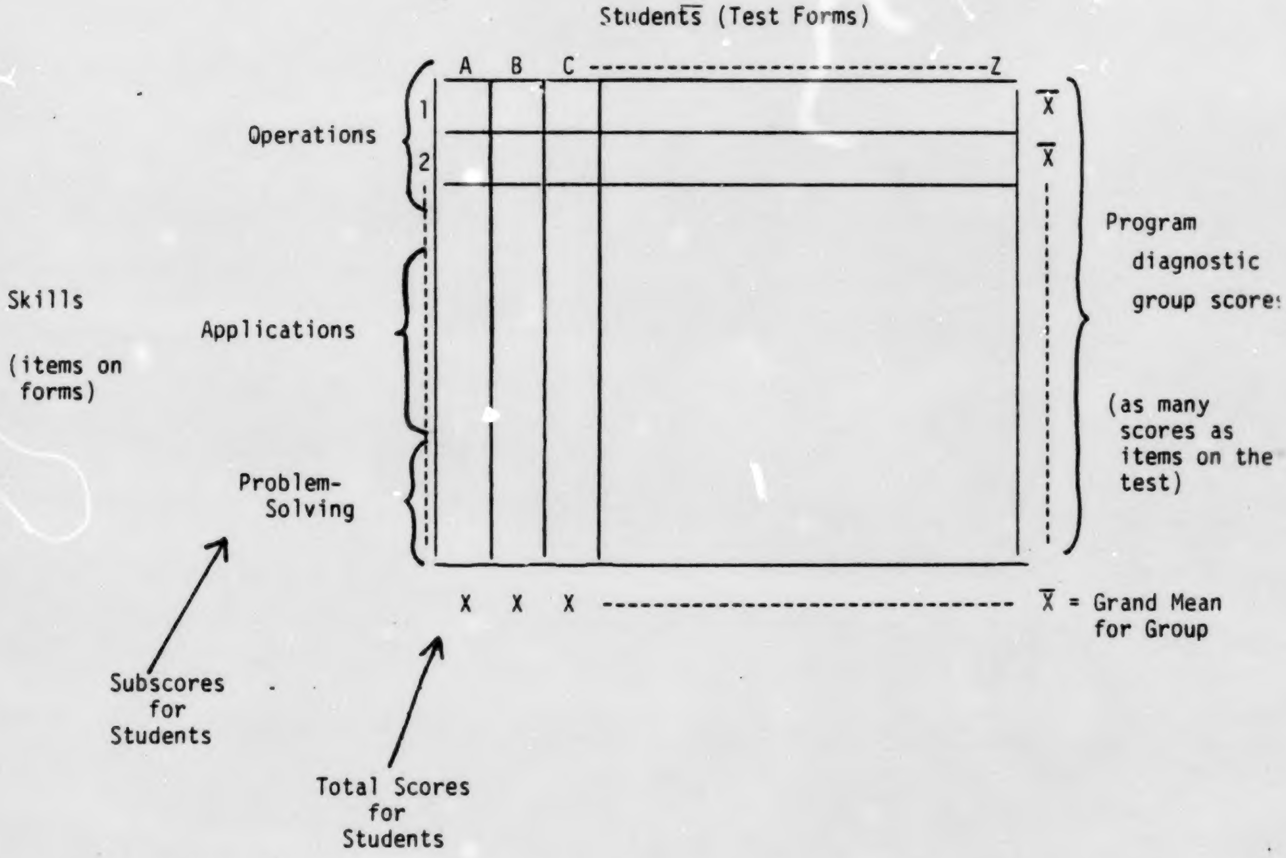
The implications of the juxtaposition of these two distinctions are important. The matrix of items by persons (see Figure 1) illustrates both distinctions and the payoff for assessment--and therefore for instruction. This matrix display shows that our traditional interest has been in (a) the overall score generated in less time and with greater reliability for a group of students and in (b) the use of group subscores to better detect the differential impact of various instructional programs, the original raison d'etre for matrix sampling. It is, however, possible with the use of new flexible, and powerful parameter estimation techniques to provide person scores summing across items independent of or at least intentionally coordinated with the development of estimates for the group as a whole, also illustrated in Figure 1. This type of design, given the general label "duplex designs" by Bock, represents the current direction of the California Assessment Program. It allows for the traditional array of multiple subscores for groups which school personnel have come to expect for curriculum and program evaluation, while still providing reliable student level scores for monitoring, selection, placement, and motivational purposes. Each student will take one of several parallel forms composed of calibrated items representing all concepts and skills. The equating of the forms, based on common skills and item response calibrating (done at the level of each skill cluster--each row in Figure 1) allows for comparable student scores; the use of different forms with different items allows for specific skill area reporting for groups. The phrase "diagnostic information" could only be applied to the type of multiple subscore information at the group level; detailed information for individual students would require additional testing--testing that obviously should be related to both the specific instructional program the student has been following and the options that are available to her or him in the future. The power and efficiency of this dual approach remain to be demonstrated, but would seem to be inevitable.

Recommendation 6: Development of content referenced reporting systems.

This history of pleas, proposals and attempts to develop a content referenced reporting system goas back at least to Reverend George Fisher, principal of Greenwich Hospital School circa 1864. He describes a "scale book" which provided examples of works of different levels of attainment and which could be used as a fixed standard against which to compare the work of individual pupils. Writings by Thurstone reveal a similar desire for a system which allows interpretation of test performance in terms of tasks which typify the skills and capabilities of students at given score

Figure 1

Illustration of a Duplex Testing Design for Mathematics

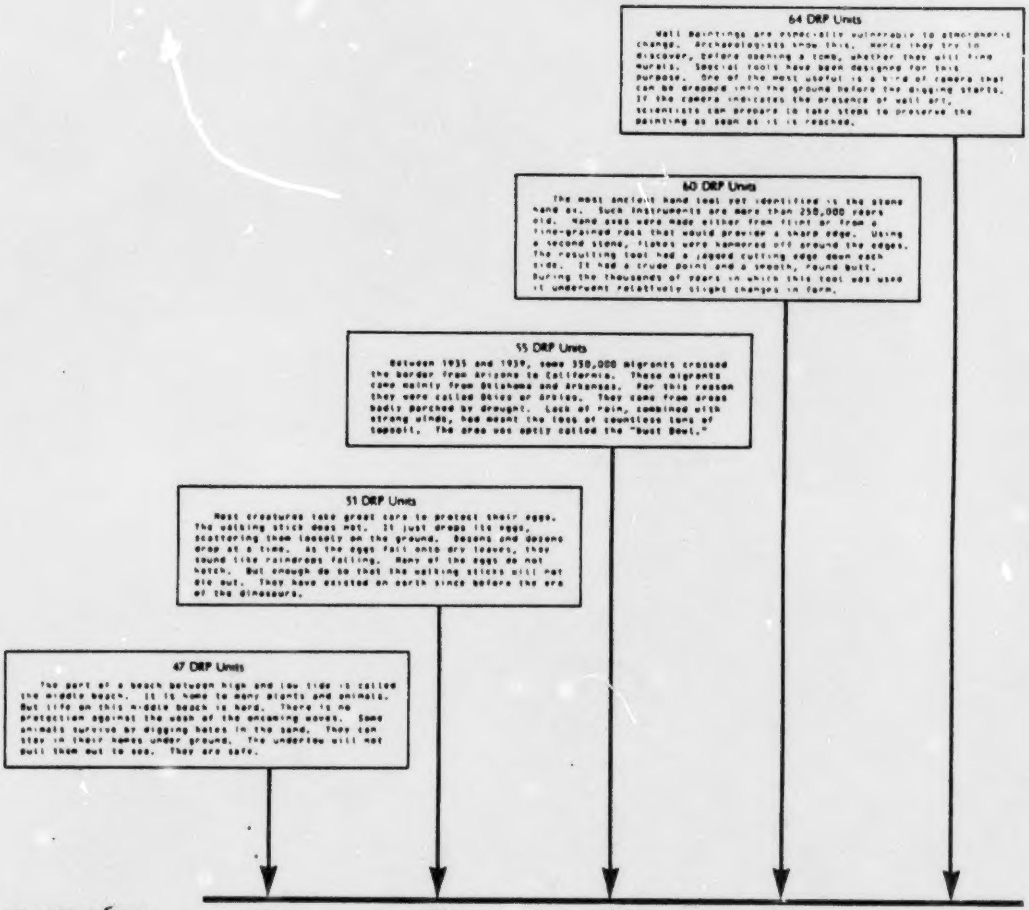

Students (Test Forms)
A
B
C
Z
1
2
Operations
Applications
Problem-Solving
Skills
(items on forms)
$\overline{X}$
$\overline{X}$
Program diagnostic group scores
(as many scores as items on the test)
X
X
X
$\overline{X}$ = Grand Mean for Group
Subscores for Students
Total Scores for Students

levels. More recently and directly, the writings of Ebel in 1962 and Glaser in 1963 call for what Darrell Bock has labeled LOCD, a linear-ordered-content-domain conception of performance. It is important to note that Glaser's seminal article which spawned criterion-referenced testing 20 years ago outined this linear idea. Glaser put it so clearly one wonders how it got lost or ignored.

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing...The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum.

Figures 2, 3 and 4 show how reading, writing, and mathematics performance could be displayed, with the assistance of item response methodology. It is obvious that the essence of criterion-referenced testing, that is, emphasis on skills rather than normative comparisons, is adequately fulfilled with this approach. Furthermore, it is consistent with reality, that is, that achievement is best represented as a continuous variable whereby the practice of identifying cut points represents an attempt to settle on an acceptable level of performance along a continuum. However, the standard setting process now openly admits of benefiting from information on the types of performance that characterize students at various score points and the proportion of students who reach each point. The unsustainable distinction between CRT and NRT approaches is erased and the information needs of proponents of both are satisfied.

It is argued that this type of reporting would assist and add credibility to test results in the eyes of the public, since it would be possible for them to more easily attach meaning to a given numerical performance, and would also assist curriculum developers and instructional design specialists in that it shows the general sequence of difficulty of skills. Not that learning is linear, sequential, and uniform, but curricular decisions and instructional design decisions can be informed in the process of determining why some tasks are more difficult than others, what role complexity plays, and what skills and knowledge structures function as propaedeutics to others in the learning hierarchy.



64 DRP Units

Wall paintings are especially vulnerable to atmospheric change. Archaeologists know this. Hence they try to discover, before opening a tomb, whether they will find murals. Special tools have been designed for this purpose. One of the most useful is a kind of camera that can be dropped into the ground before the digging starts. If the camera indicates the presence of wall art, scientists can prepare to take steps to preserve the painting as soon as it is reached.

60 DRP Units

The most ancient hand tool yet identified is the stone hand ax. Such instruments are more than 250,000 years old. Hand axes were made either from flint or from a fine-grained rock that would provide a sharp edge. Using a second stone, flakes were hammered off around the edges. The resulting tool had a jagged cutting edge down each side. It had a crude point and a smooth, round butt. During the thousands of years in which this tool was used it underwent relatively slight changes in form.

55 DRP Units

Between 1935 and 1939, some 350,000 migrants crossed the border from Arizona to California. These migrants came mainly from Oklahoma and Arkansas. For this reason they were called Okies or Arkies. They came from areas badly parched by drought. Lack of rain, combined with strong winds, had meant the loss of countless tons of topsoil. The area was aptly called the "Dust Bowl."

51 DRP Units

Most creatures take great care to protect their eggs. The walking stick does not. It just drops its eggs, scattering them loosely on the ground. Dozens and dozens drop at a time. As the eggs fall onto dry leaves, they sound like raindrops falling. Many of the eggs do not hatch. But enough do so that the walking sticks will not die out. They have existed on earth since before the era of the dinosaurs.

47 DRP Units

The part of a beach between high and low tide is called the middle beach. It is home to many plants and animals. But life on this middle beach is hard. There is no protection against the wash of the oncoming waves. Some animals survive by digging holes in the sand. They can stay in their homes under ground. The undertow will not pull them out to sea. They are safe.

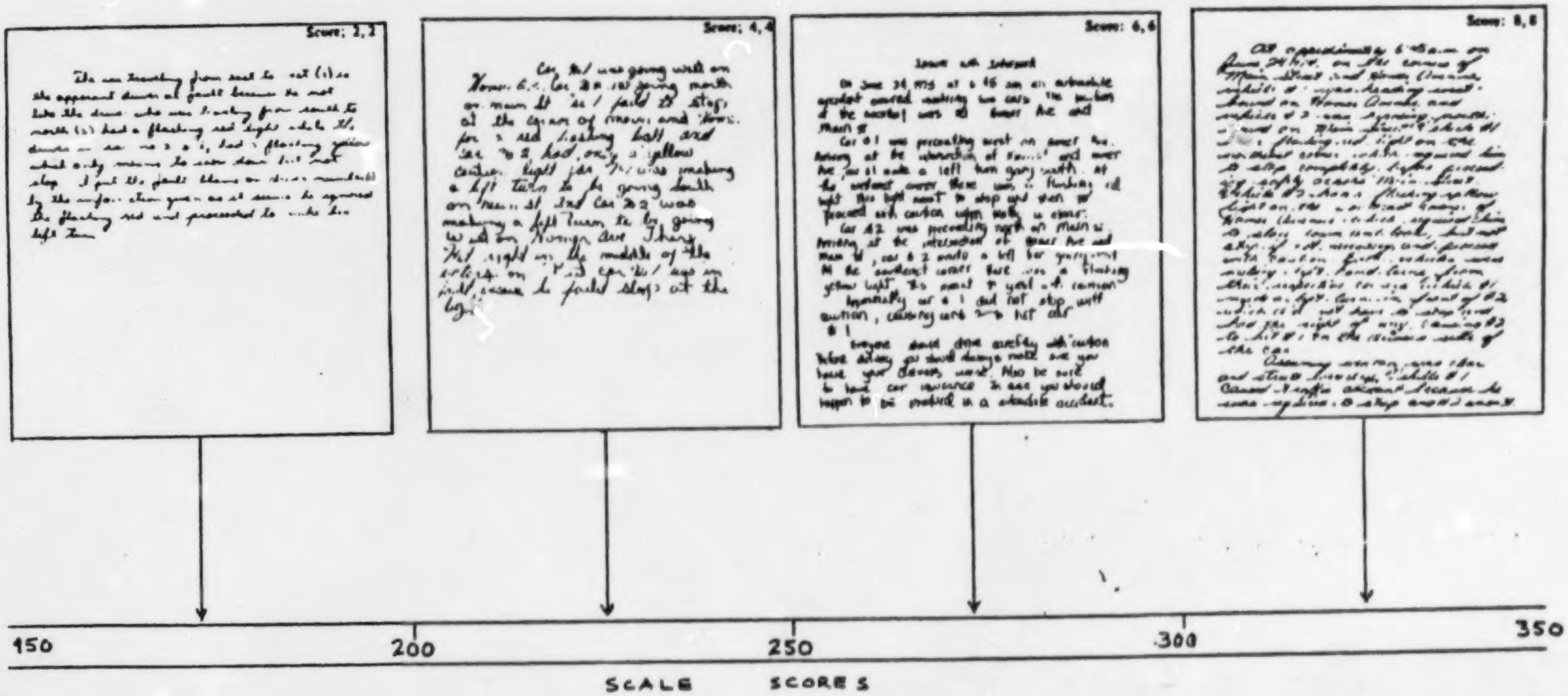
SCALED SCORE	**150**	**200**	**250**	**300**	**350**
PERCENT ABLE TO READ PASSAGE AT:					
Instructional Level	**88%**	**72%**	**51%**	**27%**	**11%**
Independent Level	**76%**	**55%**	**31%**	**15%**	**2%**

Figure 2 A prototype LOCD content-referenced report for a school.

BEST COPY AVAILABLE

Figure 3

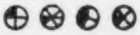
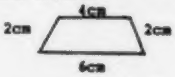
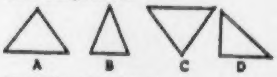
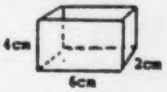
A Content-Referenced Scaled for Writing


Score: 2, 2
Score: 4, 4
Score: 6, 6
Score: 8, 8
150
200
250
300
350
SCALE SCORES

BEST COPY AVAILABLE

NAME OF STUDENT John Doe GRADE 8

Figure 4
California Assessment Program
Student Skill Area Profile

Scale Score	Skill Description	Sample Question
100	Recall basic facts	10 - 6 =
	Recognize names of numbers	702 is read as:
	Recognize place value	In 6245, the number 2 is in what place?
120	Add whole numbers	184 + 307
	Multiply whole numbers	217 x 7
140	Recognize even and odd numbers	21, 23, 28, 30, 40 How many numbers are even numbers?
160	Identify fractions	Which figure is 1/3 shaded? 
180	Find linear measures	How far is it around the figure? 
200	Word problems in place value	Paul counted paper clips by hundreds, tens, and ones. There were 205 clips in all. How many tens did he have?
220	Add/subtract decimals	7.24 - 6.83
	Divide whole numbers	$37\overline{)7281}$
240	Word problems in one step	Leah hiked 2.5 kilometers each hour. How long will it take her to hike 10 kilometers?
260	Geometric relationships	Which two figures are congruent? 
280	Find LCM and GCF	What is the least common multiple (LCM) of 24 and 6?
	Solve simple linear equation	3 x + 2 = 11; x = ?
300	Word problems involving two- or more steps	Greg needs 100 points to get extra credits in class. He received 15, 25, 30, and 16 points for the projects he has already completed. How many more points does he need?
320		
340	Compute area, volume	What is the volume of a box of the measures shown? 
360		
380	Find probability	A bag contains 2 red and 3 blue marbles. What is the probability of picking a blue marble without looking into the bag.
400		

BEST COPY AVAILABLE

Help! We Need You Now

Carl Sewell

Superintendent, Community School District 17

New York City Schools

I have been overwhelmed here with the quality of thought that has been given to the issues of testing and microcomputer technology, and I'm impressed. My role is as buyer/user of the products and processes that result from exhaustive inspections under all kinds of scientific rocks. Although I do not belittle such investigations, as a practitioner I need answers now. All I can see are problems that need resolution: budget balances, $1.3 million projected deficits, overcrowded classrooms, demands for more reduction of personnel, etc.

These are real and unromantic kinds of problems, but we practitioners need some answers not only to those kinds of questions, but to the problem of how to make whatever is happening in the classroom work better. By the time we get the issues of this conference completely figured out, we may not even be in business anymore. It is just that serious.

One speaker pointed out that the reason we have minimum competency testing and related evaluations and assessments is the erosion of public confidence in the public school system. It's not only an erosion of confidence in public school systems, it's an erosion of confidence in almost all public sector service areas, and there is growing concern about our public schools' ability to deliver--does the educational infrastructure work anymore?

I see technology through the eyes of one who needs some tools that will help me not only survive but begin to run at the head of the pack again. So I look at this whole issue of technology and testing in terms of, "Am I at the door of a fad again?" I ask myself the question, is this going to be another teaching fad that is going to constrain me even more? There's already a lack of confidence in my ability as an educator to respond to the information needs of society and the individuals within it. Or is this a whole new doorway that's going to give me and my colleagues the real freedom that we need to teach? I prefer, obviously, to think of it as the latter.

Since we're drowning out there, then, I need to concentrate and limit my comments to what I perceive as priority issues for action. I took a careful look at all of the issues that were presented here about teaching and testing and what the teacher need. There are several issues that I began to summarize and then noted that they pointed very, very much at the present technology as a potential source for resolution. There's been a lot of commentary made about "don't rush ahead, got to be careful about what we're doing, let's take a close look at some of these things." But I need something out there now. The companies that are producing materials and software for CAI-- you're right, they don't know, pedagogically, what they're doing, they really don't. I've had more conferences where people, the salespeople, get up afterwards and say to me, scratching their heads, "Hey, that's a pretty good idea. I gotta go back and talk to the folks about that one, yeah."

I'm tired of being a consultant. I'm supposed to, again, be a buyer/user. We've said mnay things about the role of the teacher in the testing process. We said that the teacher should be the major consumer of the results of testing assessment. Very good point. Because it's gotten so far away from that, that the teacher now is the one that's more or less leaning against the wall, watching the whole arena of other actors utilize testing for things that have relatively little to do with the quality of interaction between the teacher and the child, yet grossly affect whether or not the teacher's even going to be there: for example, cutting 49 teachers from the staff, or whether or not that teacher is going to have the capability to assist that child.

There is a need to make testing a relevant tool for the improvement of instruction, as opposed to just gross questions of, "Where are we? What's our benchmark as compared to other mass groups?"

There is a need to place the testing process under greater control of the teacher, the classroom teacher, as a tool, as a process to upgrade the instruction. Again, I'm trying to point at this technology.

There's a need to make greater use of test results in the formulation not just of the instructional program but of the day to day, mundane, hour to hour, minute to minute act of teaching and learning. There's a need to tie the process to the curriculum and the instructional process that implements it.

Now, we've thrown a few concepts around: textbooks driving the curriculum, tests driving the curriculum. I don't see it that way at all, and I see a way to constantly make the curriculum free. And it seems simple to me.

I think in terms of: what is it that I want the learner to learn. I think about objectives. Why am I doing this? And in response to that, I formulate goals. I form the goals into objectives and that's what determines what is in the content of the curriculum and what I want to put into that instructinal process.

What I think I really have to worry about is who's formulating the consensus. Who's involved in the consensus that this is the learning goal or objective? We have pushed the teachers out of it. I think what we have to do is bring them into that consensus and then make sure that the learning objectives are reflective of what that consensus says we need, and that that's the driving force. The test should be just a tool of the instructional process and instead it is becoming the master of it. The same thing is true of the textbook. I think we've built some false monsters that we perceive to be so real that we are reluctant to reach out, to knock them over and deal with them.

Testing is a part of the teacher's assessment techniques. We need to enhance other modes of assessment as well as testing. These two concepts: assessment, testing; I view it in the following context--I think someone already mentioned it being an overall notion of "Am I doing what I think I'm doing", and the test is a piece of that. It's just one of the ways. And I think we have to bring it back into its proper context, we must use this technology to help teachers develop and enhance some other means of assessment as well as testing.

Teachers need technical assistance and more knowledge about the preparation of what are termed internal tests or teacher-made tests. With the technology related to authoring systems and the establishment of item banks, and the very sophisticated statistical methods of developing items, it would seem to me that we are ready to bring this to the teacher along with technological delivery systems that will allow them to qualitatively enhance what they're doing. It's almost as though we're waiting and waiting and waiting and are so reluctant to jump into the waters of real world problem solving, while practioners are dying out there.

Teachers deal with enormous amounts of grunt work (paperdusting), and to me the grunt work is one of the biggest reasons why teachers don't develop very sophisticated monitoring systems for what they're doing. If you take a look at some of the paper and pencil monitoring systems that we've forced upon classroom teachers, as a result of the requirements of Title I, now Chapter I, you can easily understand teacher reaction: "I'd rather just do pupil assessment by my gut sense of 'Is this right?' or 'Is that going to be acceptable?'"

I see this technology with a tremendous capability just to take the grunt work out of instructional support processes. And that's one of the best ways to get the teachers involved in it, and to get administrators involved in it, especially when they realize "Now I can take a closer look at what I'm really doing." I see the technology as forcing the teachers to take a closer look at the quality and results of their teaching strategies and processes.

In spite of all of the psychology and the knowledge that's existent, it's not used to the degree needed on a day-to-day basis in the classroom, not from what I've seen happening there. However, when teachers realize that the computer is a system for monitoring instruction and a prescriber of what instruction should appropriately follow given feedback information, they're going to look a lot closer at the quality of the teaching-learning act.

But again, we've got to push ourselves into the water of decision making. We need to get in there and do it whether it's perfect or not. The teacher needs immediate feedback from the testing or assessment activity if it's to have maximal impact on the instructional strategy decisions made by the teacher. The present technology can provide it, if we use it. There are some systems that have been created, by no means supersophisticated, but they do help, and we do need some help in putting them together and upgrading them.

Let me just quickly cover a couple of other things. Teachers need a greater capacity to more completely (meaning in a more fine-grained way) and more accurately communicate pupil progress to pupils themselves, to parents, to other teachers, and to administrators and to the public at large. And they need to be able to do this on an individual pupil basis, or in varying types of aggregates. Present computer technology has the potential to facilitate this. Right now it's coming in the opposite way: someone on the outside starts with large aggregates, and then describes from those big aggregates, without the teacher, what's going on and the value of what's going on. I would prefer to see it starting the other way, giving the teacher the inherent control of the process. The technology will allow that fine-grained look, and the building of more accurate response in terms of "what am I doing and what's the value of it."

Testing also can serve the supervisor, which is something that hasn't been said here at all. Testing and the integration of the technology with its processing is a great tool for the manager. I've had the experience over the last two years of looking around for someone who could put together a program that would allow me to take samples on indicator skills during the year, so that I could then disaggregate that data from an individual and start aggregating it into classrooms in a given school, on a given grade, and begin to compare progress while sitting at my desk. I found a system that actually does it!

It wasn't as it was described to me by some of our colleagues. They told me I needed a massive mainframe, needed to spend thousands and thousands of dollars if I wanted to do this. I found that minicomputers are an appropriate alternative. I also found, after I added all the costs (maintenance, replacement, etc.), that over a three-year period this system would cost about $125,000. I always think of money in terms of "how many teachers is that?" It boils down to the cost of about three teachers, yet I can aid a whole school system.

Few are considering using this system as a tool for the person that has the responsibility of operating the total district instructional system. We need some help! Teachers need more information on how to utilize test results, as it's been commented upon here, for clinical decisions, instructional decisions. Even once they get the test data on how a kid did at a particular point, it's still very shaky.

In other words, I'm saying to you, this has to really be broken down. I'm not talking down about teachers, but what I'm saying is that test results have to be broken down to "so what does this have to do with the way I design my lesson plan for tomorrow." If it's not at that level, then it's not going to be used.

I made this comment, but I'll say it again. There is a need to insure that learning objectives drive curriculum and instruction, not textbooks, or publishers, or tests, and that the teachers are an integral part of the consensus that determines these learning objectives. It seems to me, then, that the microcomputer technology that we have available to us, if we consider the issues that I've raised, can capture a great deal of the findings of the CSE report and can begin to come up with some deliverable products.

The pieces are out there; let me cite a couple. For the past year and a half, we've been utilizing some systems in our school district of about 25,000 kids, serving a community of about 98% black, but not Afro-American---Afro-Caribbean, Hispanic, some others. We've been looking for software systems to give the teachers a handle on the instructional process, to raise them out of the morass of paperwork and grunt work that they have to experience on a day-to-day basis, that really resembles a wall between them and the kid.

We have looked at several systems. For example, Prescription Learning is an outfit that has put together a system that has the following kinds of components. It has a test built into the software package for diagnostic purposes. It has the capability of cataloguing all of the materials for learning in the lab, all of the "printware". It has a limited capability to add in the district or the local schools' supplies of varying printware. It then, based on the test, both the diagnostic test and the test that might be terminal after a given unit of instruction, examines the child's response and produces an instructional prescription. Clinical decision-making? A prescription. "Prescription" boils down to, here's the objective that you're trying to teach. The kid did from 0 to 100 on it, and dependent upon that response, here's what you have in your bank to go back and work on with that child.

Now, that's one level of sophistication. What I need is something that will do the following: I need something that has a whole series of diagnostic tests within it, possibly something that banks a series of diagnostic tests at varying levels going up and down, in difficulty, and going across in levels within that particular skill attainment. I also need built into this a type of authoring system that would allow me, as a teacher, to construct those diagnostic tests and those unit terminal tests. I need something that will allow me to put in all of my material, not what the publishing company wants to sell me. I need to be able to either use the computer directly, in terms of the child interacting for subsequent instruction or testing, or not use that and simply go out of the lab where the machine is and take it over to Ms. Williams's class and say, "Ms. Williams, here are the prescriptions for your children. Go for it." I also need, within this system, something that allows me to enter data without fiddling around with the keyboard or scanning devices. What I need is the capability to go from a tape right into that machine and out again. I need a machine that allows me, and it's not so much the machine, but I need the interface pieces that allow me to interface that micro with a mini or a mainframe. I understand that stuff exists. I need the stuff now, not way down the road.

I also need, when I put this lab together, the ability to use these machines instructionally because, you see, I don't have a lot of money. I can maybe only buy about 15 to 20 of these machines, and I need the ability to be able to use them on a one to one basis with pupils, and I also need to network these machines. I don't have enough money to buy four or five disk drives and a Winchester hard disk or something like a Corvus system, I can't afford that level of hardware. So I need an economical networking system as well. I have a lot of needs.

Just a few more. It's really hard to schedule a secondary school so that you can deal with the fact that Johnny is real sharp in math but he's weak in social studies, and that he's really sharp in science but he's not too good in art. In other words, I need to be able to individually schedule that child. That's a hell of a job with paper and pencil, but with a microcomputer it's a snap. I need a system that will allow me to do that.

I need one that will allow me to keep track of all these kids, too, and communicate. We talked about information systems and information facilitation, communicating. I need this thing to help me communicate the attendance, because if the child isn't sitting in the seat, I don't care what kind of instruction you have, it doesn't do any good. So, I need to communicate with the parents also. I need to integrate some systems here.

There's a need for word processing integration at this point. I need to establish data bases, not only establish them as records, but I need to be able to use them. We don't really use cumulative records now. What we do is put stuff on it, we take it, and after we fill it up then we move it to the next school and they do the same thing. We don't use the data. I want to use it to benefit the child.

Let me tell you about a wild experience that I had out in Patchaug, Long Island. I sat down at this terminal. I said, "This thing's supposed to be real sharp." The guy says, "Yeah, it's dynamite. It has an attendance package." It's 11:30 in the morning. I said, "I want to know how many kids were absent at X school." He said, "All right." Dialed it up, and here was the second absent sheet for the day. I'm at the district office. He said, "You want any other information?" I said, "Yeah, I see there's one kid here, Johnny Williams, who's been absent 70 days this semester. What's happening with Johnny?"

He says, "Well, let's take a look at the tests on him. Let's also look at the rest of the attendance." Looked at that, too. I'm still sitting there. I even knew, and this was in May, I knew that Johnny in September, on Sept. 16, was 35 minutes late to school. I then took a look at Johnny's family background, looked at his cumulative record, I looked at his test scores, and other information related to Johnny, his last grades. By the time I was finished, I had formulated a picture about that kid, thought about some action, steps that should be in place to service that child. I was then ready to pick up the phone and call that principal and that teacher, really informed. I could really supervise. I could really lead.

That system exists that I just described, and the cost is not $125,000 over three years. It looks interesting. It's not perfect, but it gives me a start.

I'm going to stop, and I'm going to leave you with what, I guess, is the practioner's perspective and point of view: I repeat the word "Help!" I'm drowning out there, in the midst of the power plays, in the midst of lessening public confidence, with the doubts about the capabilities of teachers to teach, not to mention the teachers walking through the doors who have been reduced to just warm bodies that I almost have to do a whole four-year college education all over again with, and a whole lot of other things.

Help! The technology is here, it doesn't have to be perfect, but we need to move the glowing ideas, the glowing concepts, out of the context of this kind of forum, which is not to say that this kind of forum is not vitally essential and necessary. It's the life's blood for me, the buyer, but I need you to deliver some of the goods now.

Help!

The Assessment Needs of Teachers and Administrators

Archie La Pointe

Educational Testing Service

My objective here is illustrated by a story about a young, brand-new game warden who had just shaken Mr. Watts's hand and been awarded his badge in northern California, and he was given charge of guarding one of the reservoirs to make sure the fish weren't being taken from it. And he knew that some were, but he couldn't figure out how or by whom. One morning he was walking around the reservoir and he noticed an old fisherman by the name of Clyde who was unloading heaps of fish from his rowboat. The next morning he, too, dressed as a fisherman, and rowed over to the cove and said to the old man, "Any fish in the lake?"

"Yep."

"Mind if I join you?"

"Nope."

So they climbed in the boat; they rowed out; old Clyde, when they got to the middle of the reservoir, stopped, reached into his gunnysack, took out a stick of dynamite and lit it, threw it overboard--BOOM!--and started loading all the fish that had come belly-up into the boat.

The young warden watched this and then pulled out his badge and said, "Sir, I must advise you that you're in violation of the state of California laws about about fishing in a reservoir. You have the right to remain silent. Anything you say may be held against you." And he proceeded with his dissertation. The old man looked at him, reached into his gunnysack, pulled out another stick of dynamite, lit it, handed it to the young game warden and said, "Son, you here to fish, or you here to talk?"

I'm here to talk--and shake things up a little bit. My qualification for the assignment is that of a former sixth grade schoolteacher, former frustrated parent, and some work that I've started doing with a local school system in New Jersey in trying to see how national assessment information could be of use to school districts. We have some ideas and we're trying to test them out.

So my interest in the question comes from my own problem, which is how can we make the results of the national assessment realistically pertinent and useful to classroom teachers. There's been a traditional question as to whether it should be: is that one of the functions that national assessment should serve? I can't give you a definite answer. I see it as a real challenge. I want to find an answer, and for all the reasons that have been talked about here, including Bill Coffman's statement that teachers make the curriculum; they do and in a way it's a blessing that they do, but more about that later.

So I'm going to talk about this question in relation to national assessment, connecting it to what I've heard here and to what I read in the draft of the CSE report. I have to mention that this is only one of the aspects of national assessment. The elements of the design that might interest a good many of you--the new spiraling techniques, balanced in complete blocks, our plans for scaling, for IRT scaling of the items, our addition of an elaborate teacher questionnaire, the expansion of the principals' questionnaire, the collection of an awful from children, more than has been done in the past, the intention to correlate scores and achievement scores from different subject matter areas--all these things are described in our first publication, A New Design for A New Era. But the single aspect of how NAEP test assessment information can be useful to schoolteachers is what I'd like to focus on here.

I approach this task with a fair amount of depression that comes from the experiences that William Wirtz and I had when we were asked to take a look at the National Assessment of Educational Progress. In that process we interviewed and surveyed a good many of you, a good many of all the major institutions and associations, the Elementary School Principals Association, the National School Boards Association, the NEA, the AFT, etc. I had been away from institutional education for half a dozen years when I approached this, working on human resources kinds of problems at the adult level, and I was actually shocked, and again a little depressed, to find that all the answers were institutional answers. No one was focusing on that very reality that you've been talking about this morning, which is the relationship between one teacher and one child. I came to the conclusion, which is no startling conclusion, that is the essence of what we're all about: that there are 35,000,000 kids in the K through 12 school system in the United States. There is one boy and one teacher and one girl and one teacher 17 million times. And that's what makes up the process.

Classroom teachers are aware of this and keep their minds focused on it. We seem periodically to forget this essential element. As a matter of fact, every institution as it grows faces the same problem. The New Testament writer, when he describes Jesus's interaction with the young rich man, gives us a detailed description of the content of the message of what was being taught; but the minute Jesus starts teaching 5,000 people our focus is on the quality of the food served--there are several baskets of fish and several baskets of bread left over, and the message gets lost in that remembrance.

I frankly also have a secret delight that lay people have wrenched control of the system away from us professionals. They've imposed these minimum competency tests, they've lowered our budgets, they're demanding higher standards. And maybe in all of this, the Jeffersonian belief in the common sense of the masses is going to make all this work out very well. I'm optimistic that it will. My hope is that all of us as chastened professionals are going to run out in the front of the parade again and do what we're expected to do, which is to lead and provide vision. I have a feeling that it's not going to be easy, and I have a feeling that it's going to require a fair amount of humility on our part to recognize what has been on occasion our own irrelevance. We're going to have to learn to communicate in the vernacular again, because that's what teachers talk and that's what kids understand and that's what parents seem to resonate to. And I think we have to appreciate all over again that there's as much satisfaction to be savored from teaching a young slow learner to decode the word "house" as there is in publishing another analysis of the decline of the SAT scores.

It seems to me that we have to accept that we're in the retail business. Teachers need our help. Researchers and psychologists and psychometricians and test publishers have tried too often, I think, to be in the wholesale business, in that we thought we could sell to fabricators and merchants, directors of testing, guidance counselors. people, principals, who would help teachers utilize tests. And for a host of reasons that merchant class has disappeared. They're not in the schools. So what we do has to be relevant to teachers' needs in order for them to accept it. I think that's what the teachers were telling us in the study. I frankly am delighted that teachers are as practical as they are, and that they tell us in words of one syllable what they think; therein lies our hope.

One of the things that we did learn in doing the study of national assessment is that the tests have become the standards. Educators are aware of this, judging from the work I have done with the administrators and teachers of a fairly large school district in New Jersey (approximately 15,000 students). They spent part of the time that I was with them going over the statement of competency test results and the monitoring procedures that they had undergone. As they were going over them they said, "We did well on this, we did well on this, we didn't do so well on this but we didn't know they were going to ask it, we didn't know they were going to test it," and they said ,"Next year we're going to be monitored". They, this outside force, are going to be monitoring the fifth grade. The assistant superintendent said, "Well, I can't tell you what to do as principals, but it seems to me that if were a principal and I had a terrfiic teacher in the third grade and a lousy one in the fifth grade, I'd switch for a year."

They do know, and they understand. They're very clever people. And to the extent that the tests are measuring things that are valid and that are motivating that kind of behavior, that may be the way, that may be the leverage, that will make this whole system work. There is a considerable amount of teacher intuition that grasps the facts and the reality of the situation. I don't argue with any of the findings, and I'm not surprised by a good many of them. I resonate to many comments that were made here. I told Bill Coffman that every time I listen to him, he reminds me of that other Bill--Shakespeare--who has a way of putting things into a context that makes it seem that it's all going to work out all right because the problems have been around for a long time and we've survived them. I'm sure that that's exactly what's going to happen. If any of us can contribute to making that interaction between a teacher and a youngster a little more productive, then I think that's worth the candle

Now, let me get back to my own problem and describe how we're going to address it. I see the problem of making national assessment meaningful to two million classroom teachers as a classical marketing problem. I was in the test publishing business for a while at the California Test Bureau and at SRA, and that's what gives me that perspective. And if you've got a classical marketing problem you approach it in classical ways. You identify and you describe your market. We used to say to editors and to authors and to our marketing people, sit down someday and write what your customer has for breakfast on a winter morning. In other words, get into the mindset, into the perception, into the reality of that client.

Secondly, you have to perceive their need accurately. It's so easy to come to a set of clients with a preconceived notion of what they need. And we do that because the more logical we are, the more aware we are of potential, the more we impose this logic and distort what they perceive their need to be. You have to know your product, obviously. And you must pass the test that many marketing efforts have failed: you have to match the features that are important to the client as opposed to those that are important to you either as a developer or as a thoughtful person aware of its features.

Next, you have to describe it persuasively. There are lots of people in this country who know how to do that, and they make us buy an awful lot of things. And you have to sell it enthusiastically. And finally, you have to service it faithfully. My market consists of two million teachers. They're the best that I have. They are minimally preserviced and they're inadequately inserviced. I have to respect them or get out of the business. These are not people whom I look down upon. They're people on the firing line, doing a job that I decided I didn't want anymore a long time ago. I give them credit for what they're doing and admire them for their fortitude.

The next step in marketing something that may be useful is to sell it enthusiastically. We've got some rather elaborate dissemination plans for national assessment. One of the things that we've said to ourselves over and over again is that it is not a research project. Secondly, NAEP is not a testing program. We're going to make NAEP what we think it ought to be, which is an information system. And to be a good information system, of course, it has to have the very best research base you can possibly come up with. To be a good information system, of course, it has to involve the very best assessment instruments that technology and science and the methodologies can put together.

But we've got elaborate plans for reaching the publishers who make textbooks and who make tests about what we find out, for helping state programs understand what we learn, for helping large school districts understand and use what we come up with, for reaching parents thorugh mass media, magazines, and television, and for reaching the boards of education. The boards of education are made up of 95,000 human beings who meet once or twice a month to decide what happens in 16,000 school districts. They have names and home addresses. The National Association of School Boards is going to work with us to issue two reports a year to these people.

Now, we're going to reach out to other audiences, too, that you represent. We're inheriting a data base that we're in the process of making more manageable and more useful. We're going to have computer access to that data base; you will have computer access to that data base. We're going to have 800 line numbers so people can reach us and get additional information.

Finally, if you have a product that's useful and you have a set of clients that begin to accept it and use it, it has to be serviced, and serviced faithfully. And here's where the muscle of ETS will help us. There are six regional offices across the country, one here in Los Angeles, and in each of these regional offices there will be professionals, or one professional at least, trained and ready to give workshops to teachers and to school administrators.

We have just two objectives, and they're the objectives of any good teacher. First, we want to recognize where our client is, and our client is the teacher. Our second goal is to move that person a little bit ahead in the skills they need to do their job more effectively.

My confidence is that as they develop proficiency with the instruments that we provide them, they'll find us to be more relevant and more useful. My understanding is that the essence of what I'm being paid to do for the next 5 years is to help improve the quality of that interaction between one teacher and one student. And that is the responsibility that I'm accepting.